CAN WE INCREASE INTER- AND INTRA-RATER AGREEMENT IN MODELING GENERAL MUSIC SIMILARITY?

Arthur Flexer and Taric Lallai
Austrian Research Institute for Artificial Intelligence (OFAI), Vienna, Austria
arthur.flexer@ofai.at

ABSTRACT

We present a pilot study on ways to increase inter- and intra-rater agreement in quantification of general similarity between pieces of music. By using a more controlled group of human subjects and carefully curating song material, we try to increase overall agreement between raters concerning the perceived general similarity of songs. Repeated conduction of the experiment with a two week lag shows that intra-rater agreement is higher than inter-rater agreement. Analysis of the results and interviews with test subjects suggests that the genre of songs was a major factor in judging similarity between songs. We discuss the impacts of our results on evaluation of respective machine learning models and question the validity of experiments on general music similarity.

1. INTRODUCTION

One of the successful applications of Music Information Retrieval (MIR) is the automatic recommendation of music or creation of playlists as is now commonplace and ubiquitous in music streaming services like Spotify, Deezer, Pandora or Tidal. These services often recommend music which is in some way similar to what users have been listening before. Therefore objective assessment of the quality of such services requires a quantification of similarity between recommended songs that mirrors the human perception of music similarity. Previous research [6, 8, 10, 14, 20] has shown that perception of music similarity is highly subjective with low inter-rater agreement. This is especially true for perception of *general* music similarity. Because it is not meaningful to have computational models that go beyond the level of human agreement, these levels of inter-rater agreement present a natural upper bound for any algorithmic approach [6,10,15,22,25]. To overcome this principal problem, a range of solutions have been proposed including better control of subject groups and song material, analysis of more specific aspects of music similarity, personalization of recommendations or holistic evaluation of complete MIR systems in specific use cases [6, 20]. In this paper we present a pilot study on the feasibility to improve inter-rater agreement in modeling music similarity by confining subject groups and carefully curating song material. We also report on levels of intra-rater agreement when the experiment is repeated with a two week time-lag. To the best of our knowledge, levels of intra-rater agreement have never been explored in MIR so far.

2. RELATED WORK

It seems a fundamental fact that human perception of music is highly subjective with potentially low inter-rater agreement. To give one example, if different human subjects are asked to rate the same song pairs according to their perceived similarity, only a certain amount of agreement can be expected due to a number of subjective factors [6, 20] like personal taste, musical expertise, familiarity with the music, listening history, current mood, etc. The same holds for annotation of music where different human subjects will not always agree on genre labels or other semantic tags. It was shown [23] that the performance of humans classifying songs into 19 genres ranges from modest 26% to 71% accuracy, depending on the test subject. A study [14] on transcription of chords found that annotators disagree on about 10% of harmonic annotations. A related study [10] on chord annotation showed that annotators, if given full freedom to choose chords, tend to use different chord-label vocabularies, with overlap among all annotators being less than 20%. Audio-based grounding of everyday musical terms also showed problematic results [1].

Going even further, the argument has been made [29] that music itself does not exist without the psychophysiological effect of a stimulus on humans. Therefore no such thing as an immovable 'ground' exists in the context of music, which itself is subjective, highly context-dependent and not constant. A similar conclusion has been drawn in a study on the feasibility of automatically annotating acousmatic music [9]. The same problem is also well known in general IR, where already fifty years ago it has been documented that the implicit use orientation strongly influences manual rating of retrieved items [4].

Connected to this problem, 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 users. Whenever these models are tested by new users, there will be a certain amount of disagree-

ment making it impossible that these computational models surpass the level of human agreement. This has been documented [6,8,20] 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 [5]. Since our experiment reported in Section 4 is 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 [6]. The essence of the AMS task was to have human graders evaluate 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).

Only in the year 2006 [2] every query/candidate pair in the AMS task has been evaluated by three different human graders, which makes 2006 the only year inter-rater agreement can be accessed. The average Pearson correlation between pairs of graders was found to be at the rather low level of 0.40. The same authors [6] also derived an upper bound based on ratings within the highest interval of scores, where query/candidate pairs that have been rated between 9 and 10 by one grader have received an average rating of 6.54 by the respective other two graders. This was explained to constitute 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 sounds very similar to one of the graders will on average not receive equally high scores by other graders. For AMS the upper bound has already been reached in 2009 by a number of algorithms [6].

Related results exist for the MIREX 'Music Structural Segmentation' (MSS) task, where an upper bound for MSS has been reported which is already within reach for some genres of music [6]. Additional results for music structure analysis [15, 25] and segment boundary recognition [22] exist. The level of inter-rater agreement has also been explored for melody extraction [2, 3, 18], metrical structure [17], rhythm and timbre similarity [16] as well as chord estimation [12]. An interesting new approach [11] is to use deep learning to account for different annotator styles in the task of chord labeling, essentially personalizing chord labels for individual annotators.

To the best of our knowledge, intra-rater agreement has so far not been explored in MIR, but in general IR it is a well documented fact that items are judged differently over time, even by the same people [19].

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

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

3. EXPERIMENT

Our experiment is closely 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 try to increase inter- and intra-rater agreement by introducing a number of changes. Essential differences of our experiment include: (i) a more controlled group of human graders; (ii) carefully curated song material; (iii) query/candidate pairs are not results of algorithmic recommendation but of constrained random assignment.

Participants: In selecting test subjects for this study, we were aiming at a more uniform group of persons compared to the MIREX AMS task, where participants were drawn form the larger MIR community without any restrictions and without collecting any personal data. The major selection criteria for taking part in our study was to have had musical training in the past in some way, which should give all participants a comparable background in music. We also selected participants from an essentially young age group, which resulted in all participants having been born after 1984 and their age ranging from 25 to 34 years. The sample consisted of three females and three males. All participants were personal contacts of one of the authors.

Song Material: Contrary to the MIREX AMS task with songs from nine genres, songs in this study belong to only five different genres. The genres are: (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, 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 songs can be found here.

The five genres were chosen to have small stylistic overlap. All songs originate between the 1940s and 1980s, making it more unlikely that participants are overly familiar with the music since all of them have been born after 1984. For every 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. For each genre, we always started with one stereotypical artist of the 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. The criterion for each song's degree of popularity was to have under 50.000 accesses on Spotify. Post-experiment questioning confirmed that very few songs 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 presentation in the

questionnaire, 15 seconds of a representative part of every song (usually the refrain) were chosen and normalized to 89db to control for volume effects.

Questionnaire: The study was 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 questionnaire 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 anonymously. Subsequently, the procedure of the study was explained to the participants. The subjects 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". 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 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 the sequential order being 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.

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 subjects had the possibility to leave a comment.

About two weeks after filling in the first questionnaire at time point **t1**, all subjects filled in the same questionnaire with identical randomized items a second time (time point **t2**). The test page as well as the questions about the personal experiences with music were omitted in the second round of surveys.

4. RESULTS

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 correlations between graders S1 to S6 are given in Table 2 for time t1 and in Table 3 for time t2. The correlations range from 0.59 to 0.86, with an average of 0.73 at t1 and 0.75 at t2 (see also Table 1 for an overview of results). This is considerably higher than $\rho^{AMS} = 0.40$ which has been reported for the MIREX AMS task 2006 [6]. The differences in correlation between ρ^{AMS} and correlations in our experiment are also statistically significant at both t1 ($|t| = |8.3322| > t_{95,df=16} = 2.120$) and t2 ($|t| = |8.8519| > t_{95,df=16} = 2.120$). Therefore the inter-rater agreement over the full range of scores in our experiment is increased compared to the MIREX AMS task.

	t1	t2	t1 → t2
ρ	$0.73 \pm .065$	$0.75 \pm .065$	$0.80 \pm .103$
B_{80}	67.7 ± 19.5	57.5 ± 25.6	82.1 ± 14.6
ρ^{AMS}	$0.40 \pm .027$		
B_{80}^{AMS}	61.65 ± 27.0		

Table 1. Overview of results for time points t1, t2 and between t1 and t2 (t1 → t2). Shown are average correlations ρ and upper bounds $B_{80} \pm$ standard deviations, also for MIREX AMS task (last two lines).

Next we explore the level of agreement for specific intervals of scores. We 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. We therefore explore how human graders rate pairs of songs which another human grader rated at a specific level of similarity. The average results across all raters and for intervals v ranging from $[0, 1], (1, 2]$... to $(9, 10]$ are plotted in Figure 1 for t1 and in Figure 2 for t2. It is evident that there is a certain deviation from the theoretical perfect agreement which is 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. On the other end of the spectrum, query/candidate pairs rated as being not similar at all (score between 0 and 10) receive average scores of 17.4 (t2) and 16.1 (t2) by the respective other raters.

In a previous study [6] an upper bound B^{AMS} has been reported based on scores within the highest interval. 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 such scores for t1 and 47 for t2. This upper bound B_{80} is 67.7 for t1 and 57.5 for t2 (see Table 1). If we multiply scores from AMS 2006 by ten for better comparability and compute a similar upper bound B_{80}^{AMS} this 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 Pearson correlation ρ between t1 and t2 for the 75 pairs of query/candidate songs for graders S1 to S6 are given in Table 4. The correla-

	S1	S2	S3	S4	S5	S6
S1	1.00	0.77	0.74	0.72	0.74	0.82
S2		1.00	0.72	0.75	0.62	0.83
S3			1.00	0.70	0.67	0.76
S4				1.00	0.64	0.80
S5					1.00	0.64
S6						1.00

Table 2. Inter-rater correlation at time **t1**, with mean $0.73 \pm .065$ standard deviation.

	S1	S2	S3	S4	S5	S6
S1	1.00	0.79	0.73	0.77	0.74	0.83
S2		1.00	0.73	0.74	0.75	0.86
S3			1.00	0.69	0.69	0.80
S4				1.00	0.59	0.79
S5					1.00	0.73
S6						1.00

Table 3. Inter-rater correlation at time **t2**, with mean $0.75 \pm .065$ standard deviation.

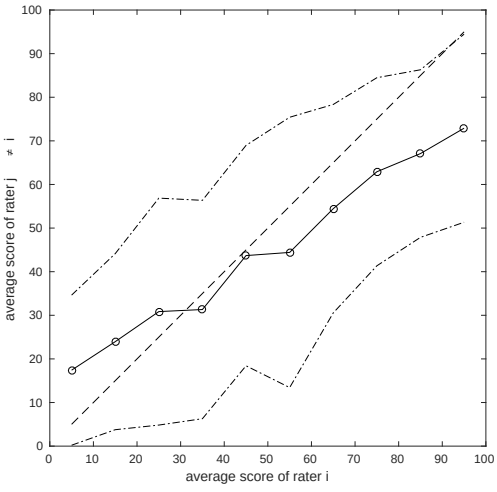

Figure 1. Average score inter-rater agreement for different intervals of scores (solid line) $\pm$ one standard deviation (dash-dot lines) at time **t1**. Dashed line indicates theoretical perfect agreement.

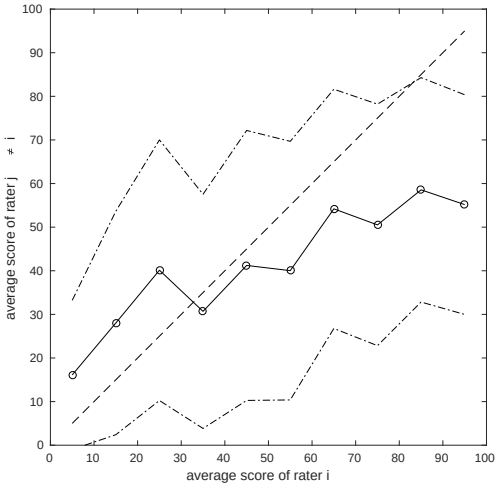

Figure 2. Average score inter-rater agreement for different intervals of scores (solid line) $\pm$ one standard deviation (dash-dot lines) at time **t2**. Dashed line indicates theoretical perfect agreement.

S1	S2	S3	S4	S5	S6
0.81	0.85	0.75	0.81	0.64	0.95

Table 4. Intra-rater correlation between times **t1** and **t2**, with mean $0.80 \pm .103$ standard deviation.

tions 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 1). The differences between inter-agreement correlations and intra-agreement correlation are however not statistically significant, neither for t1 ($|t| = |-1.9742| < t_{95,df=19} = 2.093$) nor for t2 ($|t| = |-1.3682| < t_{95,df=19} = 2.093$).

Similar to what we did for inter-rater agreement, we also plotted the average score of a rater i for all query/candidate pairs, which he or she rated within a certain interval of scores v at time t1, versus the average scores achieved by the same rater i for the same query/candidate pairs at time t2 in Figure 3. Compared to Figures 1 and 2, we see better agreement within persons between t1 and t2, with intra-agreement being very close to theoretical perfect agreement (dashed line) for scores ranging from 0 to 50, but also from 90 to 100.

We also computed an upper bound B^{80} based on ratings within the interval $(80, 100]$, where query/candidate pairs that have been rated between 80 and 100 by a grader i at t1 have 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 ($|t| = |-4.2537| > t_{95,df=220} = 1.960$) and t2 ($|t| = |-5.7121| > t_{95,df=19} = 1.960$) Therefore we conclude that the upper bound within participants measured with a two week time lag is higher then the upper bound based on inter-rater agreement.

Because a number of participants in this study com-

	Soul	Bebop	High Energy	70s Rock	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
70s Rock	17.4	-	20.9	**48.2**	11.0
Rocksteady	35.0	-	23.3	13.3	**66.1**

Table 5. Genre confusion matrix at time **t1**, shown are average scores per genre combination.

	Soul	Bebop	High Energy	70s Rock	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
70s Rock	18.6	-	16.8	**50.6**	11.0
Rocksteady	26.1	-	15.9	8.8	**62.5**

Table 6. Genre confusion matrix at time **t2**, shown are average scores per genre combination.

mented 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 Tables 5 and 6 we present genre confusion matrices at times t1 and t2. Just to give one example, the first entry in the first line in Table 5 shows that whenever both query and candidate song were from genre 'Soul', on average such pairs have been 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 confusion 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 then all off-diagonal entries. This 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. '70s Rock'.

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 correct 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.

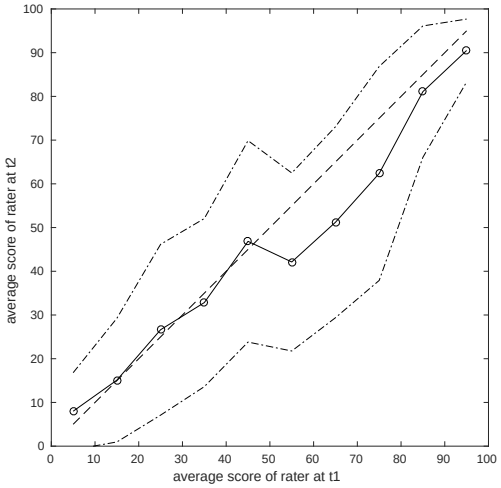

Figure 3. Average score intra-rater agreement for different intervals of scores (solid line) $\pm$ one standard deviation (dash-dot lines) between times **t1** and **t2**. Dashed line indicates theoretical perfect agreement.

5. DISCUSSION

Our principal reason for conducting this research and experiment is the quest to raise the level of rater agreement when judging music similarity. 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 has been reached already in 2009 and has not been outperformed ever since [6]. As a matter of fact, the respective MIREX task (AMS) has been dormant since 2015. In an attempt to improve the AMS task, we have conducted an experiment with a more controlled group of participants and with more carefully curated song material.

The overall inter-rater agreement measured via correla-

tion could indeed be increased 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. Therefore our new version of the AMS task based on more controlled participants and better curated song material is not better suited to measure differences between algorithms that have already reached the AMS upper bound.

One proposal 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. Certainly this is what many commercial services do by offering individual recommendations tailored to their users. This of course brings us to the question how stable assessment of music similarity is within persons, i.e. when the same persons have to judge music similarity repeatedly. The result concerning this intra-rater agreement in our experiment is divided. On the one hand the overall agreement as measured via correlation could not be improved, or 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.

For future efforts to improve on the original AMS task design, one should probably rethink which songs from what genre to use. In our experiment, genres were so distinct that at least some participants used membership to a certain genre as the main criterion when assessing similarity between songs. This might impact generalizability of our results when judging music similarity within individual genres. It is also not completely clear, what contribution to improvements our more controlled group of subjects had. We are however convinced that the more uniform group of subjects with a certain 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.

One should also point out that this is a pilot study with only six participants. A higher number of test subjects might have allowed for more statistically significant results, e.g. concerning differences between inter and intra-rater correlations. Future work should also explore alternatives to Pearson correlation like generalizations of Kappa measures to the interval scale, which take into account the possibility of rater agreement occurring by chance (e.g. intra-class-correlation [24]). Another open question is whether a time lag of more than two weeks might change results on intra-rater agreement.

A possible route to further improvements is to ask a more specific question than just assessing general music similarity, as in the original AMS task and in the experiment reported in this paper. Criticism of this unclear abstract notion of general music similarity brings us to the concept of 'validity' of our experiment. A valid experiment is an experiment that actually measures what the experimenter intended to measure (see e.g. [27] or [28] for a discussion in relation to MIR). Precisely this intention of the experimenter in the original MIREX AMS task is completely 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 [20]. 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. It has also been noted before that evaluation of abstract music similarity without reference to a specific usage scenario is not very meaningful [7, 21]. It is therefore our belief that the intention of a music similarity experiment can only be made clearer if it will be tied 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 [26] in MIR. Previous reviews [13] of user studies in MIR could serve as valuable input for formalization of the use cases.

6. CONCLUSION

We have presented a pilot study aimed at improving experiments to measure general music similarity. By using a more controlled group of subjects and music material from more well defined genres, we were able to improve overall inter-rater agreement but 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. We also discussed the doubtful validity of experiments on general music similarity making it clear that definition of a specific use case might be necessary for conduction of a truly valid experiment. The fact that MIR needs to care much more about the proper design of its experiments is also the main insight going beyond the scope of this paper. Although a small but growing number of publications concerning design and evaluation of MIR experiments exists, they have so far not been able to change the research culture of MIR as a whole.

Acknowledgements: This work was supported by the Austrian Science Fund (FWF P31988) and the Vienna Science and Technology Fund (WWTF MA14-018).

7. REFERENCES

[1] Aucouturier J.J.: Sounds like teen spirit: Computational insights into the grounding of everyday musical terms, in Minett J.W., Wang W. S-Y. (eds.): Language, Evolution and the Brain, City University of Hong Kong Press, pp. 35-64, 2009.

[2] Balke S., Driedger J., Abeßer J., Dittmar C., Müller, M.: Towards Evaluating Multiple Predominant Melody

Annotations in Jazz Recordings, in Proc. of the 17th Int. Society for Music Information Retrieval Conference, pp. 246-252, 2016.

[3] Bosch J., Gómez E.: Melody extraction in symphonic classical music: a comparative study of mutual agreement between humans and algorithms, in Proc. of the 9th Conference on Interdisciplinary Musicology, 2014.

[4] Cuadra C., Katter R.: Opening the black box of "relevance", J. of Documentation, 23(4), 291-303, 1967.

[5] Downie J.S.: The Music Information Retrieval Evaluation eXchange (MIREX), D-Lib Magazine, Volume 12, Number 12, 2006.

[6] Flexer A., Grill T.: The Problem of Limited Inter-rater Agreement in Modelling Music Similarity, J. of New Music Research, Vol. 45, No. 3, pp. 239-251, 2016.

[7] Hu, X., Liu J.: Evaluation of music information retrieval: Towards a user-centered approach, in Proceedings of the 4th Workshop on Huan-Computer Interaction and Information Retrieval, pp. 111-114, 2010.

[8] Jones M.C., Downie J.S., Ehmann A.F.: Human similarity judgments: Implications for the design of formal evaluations, in Proc. of the 8th Int. Conference on Music Information Retrieval, pp. 539-542, 2007.

[9] Klien V., Grill T., Flexer, A.: On automated annotation of acousmatic music, J. of New Music Research, 41(2), 153-173, 2012.

[10] Koops H.V.: Computational modelling of variance in musical harmony, PhD thesis, University of Utrecht, The Netherlands, 2019.

[11] Koops H.V., de Haas W.B., Bransen J., Volk A.: Automatic chord label personalization through deep learning of shared harmonic interval profiles, Neural Computing and Applications, 2018.

[12] Koops H.V., de Haas W.B., Burgoyne J.A., Bransen J., Kent-Muller A., Volk A.: Annotator subjectivity in harmony annotations of popular music, J. of New Music Research, 48:3, 232-252, 2019.

[13] Lee J.H., Cunningham S.J.: Toward an understanding of the history and impact of user studies in music information retrieval, J. of Intelligent Information Systems, 41, pp. 499-521, 2013.

[14] Ni Y., McVicar M., Santos-Rodriguez R., De Bie T.: Understanding effects of subjectivity in measuring chord estimation accuracy, IEEE Transactions on Audio, Speech and Language Processing, 21(12):2607-2615, 2013.

[15] Nieto O., Farbood M.M., Jehan T., and Bello, J.P.: Perceptual analysis of the f-measure for evaluating section boundaries in music, in Proc. of the 15th Int. Society for Music Information Retrieval Conference, pp. 265-270, 2014.

[16] Panteli M., Rocha B., Bogaards N., Honingh A.: A model for rhythm and timbre similarity in electronic dance music, Musicae Scientiae, 21(3), 338-361, 2017.

[17] Quinton E., Harte C., Sandler M.: Extraction of metrical structure from music recordings, in Proc. of the 18th Int. Conference on Digital Audio Effects, 2015.

[18] Salamon J., Gómez E., Ellis D.P.W., Richard G.: Melody extraction from polyphonic music signals: Approaches, applications, and challenges, IEEE Signal Processing Magazine, 31(2):118-134, 2014.

[19] Schamber L.: Relevance and Information Behavior, Annual Review of Information Science and Technology, 29:3-48, 1994.

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

[21] Serrá X., Magas M., Benetos E., Chudy M., Dixon S., Flexer A., Gomez E., Gouyon F., Herrera P., Jorda S., Paytuvi O., Peeters G., Schlüter J., Vinet H., Widmer G., *Roadmap for Music Information ReSearch*, Peeters G. (ed.), 2013, Creative Commons BY-NC-ND 3.0 license, ISBN: 978-2-9540351-1-6.

[22] Serrà J., Müller M., Grosche P., Arcos J.L.: Unsupervised music structure annotation by time series structure features and segment similarity, IEEE Transactions on Multimedia, 16(5): 1229-1240, 2014.

[23] Seyerlehner K., Widmer G., Knees P.: A Comparison of Human, Automatic and Collaborative Music Genre Classification and User Centric Evaluation of Genre Classification Systems, in Proc. of the 8th Int. Workshop on Adaptive Multimedia Retrieval, pp. 118-131, 2010.

[24] Shrout P.E., Fleiss J.L.: Intraclass correlation: Uses in assessing rater reliability, Psychological Bulletin, 86, 420-428, 1979.

[25] Smith J.B., Chew E.: A meta-analysis of the mirex structure segmentation task, in Proceddings of the 14th International Society for Music Information Retrieval Conference, pp. 45-47, 2013.

[26] Sturm B.L.: The state of the art ten years after a state of the art: Future research in music information retrieval, J. of New Music Research, 43 (2), pp. 147-172, 2014.

[27] Trochim, W.: The Research Methods Knowledge Base, 2nd edn, Atomic Dog Publishing, Cincinnati, OH, 2000.

[28] Urbano J., Schedl M., Serra X.: Evaluation in music information retrieval, J. of Intelligent Information Systems, 41 (3), pp. 345-369, 2013.

[29] Wiggins G.: Semantic Gap?? Schemantic Schmap!! Methodological Considerations in the Scientific Study of Music, in Proc. of the 11th IEEE Int. Symposium on Multimedia, pp. 477-482, 2009.