ON INTER-RATER AGREEMENT IN AUDIO MUSIC SIMILARITY

Arthur Flexer
Austrian Research Institute for Artificial Intelligence (OFAI)
Freyung 6/6, Vienna, Austria
arthur.flexer@ofai.at

ABSTRACT

One of the central tasks in the annual MIREX evaluation campaign is the ”Audio Music Similarity and Retrieval (AMS)” task. Songs which are ranked as being highly similar by algorithms are evaluated by human graders as to how similar they are according to their subjective judgment. By analyzing results from the AMS tasks of the years 2006 to 2013 we demonstrate that: (i) due to low inter-rater agreement there exists an upper bound of performance in terms of subjective gradings; (ii) this upper bound has already been achieved by participating algorithms in 2009 and not been surpassed since then. Based on this sobering result we discuss ways to improve future evaluations of audio music similarity.

1. INTRODUCTION

Probably the most important concept in Music Information Retrieval (MIR) is that of *music similarity*. Proper modeling of music similarity is at the heart of every application allowing automatic organization and processing of music databases. Consequently, the ”Audio Music Similarity and Retrieval (AMS)” task has been part of the annual ”Music Information Retrieval Evaluation eXchange” (MIREX [1]) [2] since 2006. MIREX is an annual evaluation campaign for MIR algorithms allowing for a fair comparison in standardized settings in a range of different tasks. As such it has been of great value for the MIR community and an important driving force of research and progress within the community. The essence of the AMS task is to have human graders evaluate pairs of query/candidate songs. The query songs are randomly chosen from a test database and the candidate songs are recommendations automatically computed by participating algorithms. The human graders rate 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).

It is precisely this general notion of ”sounding similar” which is the central point of criticism in this paper. A recent survey article on the ”neglected user in music information retrieval research” [13] has made the important argument that users apply very different, individual notions of similarity when assessing the output of music retrieval systems. It seems evident that music similarity is a multi-dimensional notion including timbre, melody, harmony, tempo, rhythm, lyrics, mood, etc. Nevertheless most studies exploring music similarity within the field of MIR, which are actually using human listening tests, are restricted to overall similarity judgments (see e.g. [10] or [11, p. 82]) thereby potentially blurring the many important dimensions of musical expression. There is very little work on what actually are important dimensions for humans when judging music similarity (see e.g. [19]).

This paper therefore presents a meta analysis of all MIREX AMS tasks from 2006 to 2013 thereby demonstrating that: (i) there is a low inter-rater agreement due to the coarse concept of music similarity; (ii) as a consequence there exists an upper bound of performance that can be achieved by algorithmic approaches to music similarity; (iii) this upper bound has already been achieved years ago and not surpassed since then. Our analysis is concluded by making recommendations on how to improve future work on evaluating audio music similarity.

2. RELATED WORK

In our review on related work we focus on papers directly discussing results of the AMS task thereby adressing the problem of evaluation of audio music similarity.

After the first implementation of the AMS task in 2006, a meta evaluation of what has been achieved has been published [8]. Contrary to all subsequent editions of the AMS task, in 2006 each query/candidate pair was evaluated by three different human graders. Most of the study is concerned with the inter-rater agreement of the BROAD scores of the AMS task as well as the ”Symbolic Melodic Similarity (SMS)” task, which followed the same evaluation protocol. To access the amount of agreement, the authors use Fleiss’s Kappa [4] which ranges between 0 (no agreement) and 1 (perfect agreement). Raters in the AMS task achieved a Kappa of 0.21 for the BROAD task, which can be seen as a ”fair” level of agreement. Such a ”fair” level of agreement [9] is given if the Kappa result is between 0.21 and 0.40, therefore positioning the

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

BROAD result at the very low end of the range. Agreement in SMS is higher (Kappa of 0.37), which is attributed to the fact that the AMS task is ”less well-defined” since graders are only informed that ”works should sound similar” [8]. The authors also note that the FINE scores for query/candidate pairs, which have been judged as ”somewhat similar”, show more variance then the one judged as ”very” or ”not” similar. One of the recommendations of the authors is that ”evaluating more queries and more candidates per query would more greatly benefit algorithm developers” [8], but also that a similar analysis of the FINE scores is also necessary.

For the AMS task 2006, the distribution of differences between FINE scores of raters judging the same query/candidate pair has already been analysed [13]. For over 50%, the difference between rater FINE scores is larger than 20. The authors also note that this is very problematic since the difference between the best and worst AMS 2012 systems was just 17.

In yet another analysis of the AMS task 2006, it has been reported [20] that a range of so-called ”objective” measures of audio similarity are highly correlated with subjective ratings by human graders. These objective measures are based on genre information, which can be used to automatically rank different algorithms producing lists of supposedly similar songs. If the genre information of the query and candidate songs are the same, a high degree of audio similarity is achieved since songs within a genre are supposed to be more similar than songs from different genres. It has therefore been argued that, at least for large-scale evaluations, these objective measures can replace human evaluation [20]. However, this is still a matter of controversy within the music information retrieval community, see e.g. [16] for a recent and very outspoken criticism of this position.

A meta study of the 2011 AMS task explored the connection between statistical significance of reported results and how this relates to actual user satisfaction in a more realistic music recommendation setting [17]. The authors made the fundamental clarification that the fact of observing statistically significant differences is not sufficient. More important is whether this difference is noticeable and important to actual users interacting with the systems. Whereas a statistically significant difference can alway be achieved by enlarging the sample size (i.e. the number of query/candidate pairs), the observed difference can nevertheless be so small that it is of no importance to users. Through a crowd-sourced user evaluation, the authors are able to show that there exists an upper bound of user satisfaction with music recommendation systems of about 80%. More concretely, in their user evaluation the highest percentage of users agreeing that two systems ”are equally good” never exceeded 80%. This upper bound cannot be surpassed since there will always be users that disagree concerning the quality of music recommendations. In addition the authors are able to demonstrate that differences in FINE scores, which are statistically significant, are so small that they make no practical difference for users.

3. DATA

For our meta analysis of audio music similarity (AMS) we use the data from the “Audio Music Similarity and Retrieval” tasks from 2006 to 2013 [2] within the annual MIREX [2] evaluation campaign for MIR algorithms.

For the AMS 2006 task, 5000 songs were chosen from the so-called ”uspop”, ”uscrap” and ”cover song” collections. Each of the participating 6 system then returned a 5000x5000 AMS distance matrix. From the complete set of 5000 songs, 60 songs were randomly selected as queries and the first 5 most highly ranked songs out of the 5000 were extracted for each query and each of the 6 systems (according to the respective distance matrices). These 5 most highly ranked songs were always obtained after filtering out the query itself, results from the same artist (i.e. a so-called artist filer was employed [5]) and members of the cover song collection (since this was essentially a separate task run together with the AMS task). The distribution for the 60 chosen random songs is highly 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 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 labeled as Rock/Pop, while a further 25% are Rap & Hip-Hop)” [3]. For each query song, the returned results (candidates) from all participating systems were evaluated by human graders. For each individual query/candidate pair, three different human graders provided both a FINE score (from 0 (failure) to 10 (perfection)) and a BROAD score (not similar, somewhat similar, very similar) indicating how similar the songs are in their opinion. This altogether gives $6 \times 60 \times 5 \times 3 = 5400$ human FINE and BROAD gradings. Please note that since some of the query/candidate pairs are identical for some algorithms (i.e. different algorithms returned identical candidates) and since such identical pairs were not graded repeatedly, the actual number of different FINE and BROAD gradings is somewhat smaller.

Starting with the AMS task 2007, a number of small changes to the overall procedure was introduced. Each participating algorithm was given 7000 songs chosen from the ”uspop”, ”uscrap” and ”american” ”classical” and ”sundry” collections. Therefore there is only a partial overlap in music collections (”uspop” and ”uscrap”) compared to AMS 2006. From now on 30 second clips instead of the full songs were being used both as input to the algorithms and as listening material for the human graders. For the subjective evaluation of music similarity, from now on 100 query songs were randomly chosen representing the 10 genres found in the database (i.e., 10 queries per genre). The whole database consists of songs from equally sized genre groups: BAROQUE, COUNTRY, EDANCE,

[2] The results and details can be found at: http://www.music-ir.org/mirex/wiki/MIREX_HOME

[3] This is stated in the 2006 MIREX AMS results: http://www.music-ir.org/mirex/wiki/2006:Audio_Music_Similarity_and_Retrieval_Results

JAZZ, METAL, RAPHIPHOP, ROCKROLL, ROMANTIC, BLUES, CLASSICAL. Therefore there is only a partial overlap of genres compared to AMS 2006 (COUNTRY, EDANCE, JAZZ, RAPHIPHOP, ROCKROLL). As with AMS 2006, the 5 most highly ranked songs were then returned per query as candidates (after filtering for the query song and songs from the same artist). For AMS tasks 2012 and 2013, 50 instead of 100 query songs were chosen and 10 instead of 5 most highly ranked songs returned as candidates.

Probably the one most important change to the AMS 2006 task is the fact that from now on every query/candidate pair was only being evaluated by a single user. Therefore the degree of inter-rater agreement cannot be analysed anymore. For every AMS task, the subjective evaluation therefore results in $a \times 100 \times 5$ human FINE and BROAD gradings, with a being the number of participating algorithms, 100 the number of query songs and 5 the number of candidate songs. For AMS 2012 and 2013 this changed to $a \times 50 \times 10$, which yields the same overall number. These changes are documented on the respective MIREX websites, but also in a MIREX review article covering all tasks of the campaign [3]. For AMS 2007 and 2009, the FINE scores range from 0 to 10, from AMS 2010 onwards from 0 to 100. There was no AMS task in MIREX 2008.

4. RESULTS

In our meta analysis of the AMS tasks from years 2006 to 2013, we will focus on the FINE scores of the subjective evaluation conducted by the human graders. The reason is that the FINE scores provide more information than the BROAD scores which only allow for three categorical values. It has also been customary for the presentation of AMS results to mainly compare average FINE scores for the participating algorithms.

4.1 Analysis of inter-rater agreement

Our first analysis is concerned with the degree of inter-rater agreement achieved in the AMS task 2006, which is the only year every query/candidate pair has been evaluated by three different human graders. Previous analysis of AMS results has concentrated on BROAD scores and used Fleiss's Kappa as a measure of agreement (see Section 2). Since the Kappa measure is only defined for the categorical scale, we use the Pearson correlation ρ between FINE scores of pairs of graders. As can be seen in Table 1, the average correlations range from 0.37 to 0.43. Taking the square of the observed values of ρ, we can see that only about 14 to 18 percent of the variance of FINE scores observed in one grader can be explained by the values observed for the respective other grader (see e.g. [1] on ρ^2 measures). Therefore, this is the first indication that agreement between raters in the AMS task is rather low.

	grader1	grader2	grader3
grader1	1.00	0.43	0.37
grader2		1.00	0.40
grader3			1.00

Table 1. Correlation of FINE scores between pairs of human graders.

	grader1	grader2	grader3
grader1	9.57	6.66	5.99
grader2	6.60	9.55	6.67
grader3	6.62	6.87	9.69

Table 2. Pairwise inter-rater agreement for FINE scores from interval $v = [9, 10]$.

Next we plotted the average FINE score of a rater i for all query/candidate pairs, which he or she rated within a certain interval of FINE scores v, versus the average FINE scores achieved by the other two 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. It is evident that there is a considerable deviation from the theoretical perfect agreement which is indicated as a dashed line. Pairs of query/candidate songs which are rated as being very similar (FINE score between 9 and 10) by one grader are on average only rated at around 6.5 by the two other raters. On the other end of the spectrum, query/candidate pairs rated as being not similar at all (FINE score between 0 and 1) receive average FINE scores of almost 3 by the respective other raters. The degree of inter-rater agreement for pairs of raters at the interval $v = [9, 10]$ is given in Table 2. There are 333 pairs of songs which have been rated within this interval. The main diagonal gives the average rating one grader gave to pairs of songs in the interval $v = [9, 10]$. The off-diagonal entries show the level of agreement between different raters. As an example, query/candidate pairs that have been rated between 9 and 10 by *grader1* have received an average rating of 6.66 by *grader2*. The average of these pairwise inter-rater agreements given in Table 2 is 6.54 and is an upper bound for the average FINE scores of the AMS task 2006. This upper bound is the maximum of average FINE scores that can be achieved within such an evaluation setting. This upper bound is due to the fact that there is 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.

The average FINE score achieved by the best participating system in AMS 2006 (algorithm EP) is 4.30 ± 8.8 (mean ± variance). The average upper bound inter-rater grading is 6.54 ± 6.96. The difference between the best FINE scores achieved by the system EP and the upper bound is significant according to a t-test: $|t| = |-12.0612| > t_{95,df=1231} = 1.96$ (confidence level of 95%, degrees of freedom = 1231). We can therefore conclude that for the AMS 2006 task, the upper bound on the av-

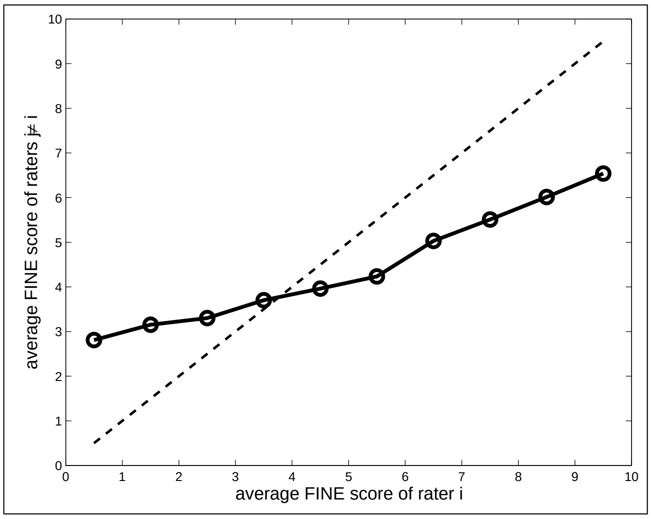

Figure 1. Average FINE score inter-rater agreement for different intervals of FINE scores (solid line). Dashed line indicates theoretical perfect agreement.

erage FINE score had not yet been reached and that there still was room for improvement for future editions of the AMS task.

4.2 Comparison to the upper bound

We will now compare the performance of the respective best participating systems in AMS 2007, 2009 to 2013 to the upper bound of average FINE scores we have retrieved in Section 4.1. This upper bound that can possibly be achieved due to the low inter-rater agreement results from the analysis of the AMS 2006 task. Although the whole evaluation protocol in all AMS tasks over the years is almost identical, AMS 2006 did use a song database that is only overlapping with that of subsequent years. It is therefore of course debatable how strictly the upper bound from AMS 2006 applies to the AMS results of later years. As outlined in Section 3, AMS 2006 has a genre distribution that is skewed to about 50% of rock music whereas all other AMS databases consist of equal amounts of songs from 10 genres. One could make the argument that in general songs from the same genre are being rated as being more similar than songs from different genres. As a consequence, agreement of raters for query/candidate pairs from identical genres might also be higher. Therefore inter-rater agreement within such a more homogeneous database should be higher than in a more diverse database and it can be expected that an upper bound of inter-rater agreement for AMS 2007 to 2013 is even lower than the one we obtained in Section 4.1. Of course this line of argument is somewhat speculative and needs to be further investigated.

In Figure 2 we have plotted the average FINE score of the highest performing participants of AMS tasks 2007, 2009 to 2013. These highest performing participants are the ones that achieved the highest average FINE scores in the respective years. In terms of statistical significance, the performance of these top algorithms is often at the same level as a number of other systems. We have also plotted the upper bound (dashed line) and a 95% confidence interval (dot-dashed lines). As can be seen the performance peaked in the year 2009 where the average FINE score reached the confidence interval. Average FINE scores in all other years are always a little lower. In Table 3 we show the results of a number of t-tests always comparing the performance to the upper bound. Table 3 gives the AMS year, the abbreviated name of the winning entry, the mean performance, its variance and the resulting t-value (with 831 degrees of freedom and 95% confidence). Only the best entry from year 2009 (PS2) reaches the performance of the upper bound, the best entries from all other years are statistically significant below the upper bound (critical value for all t-tests is again 1.96).

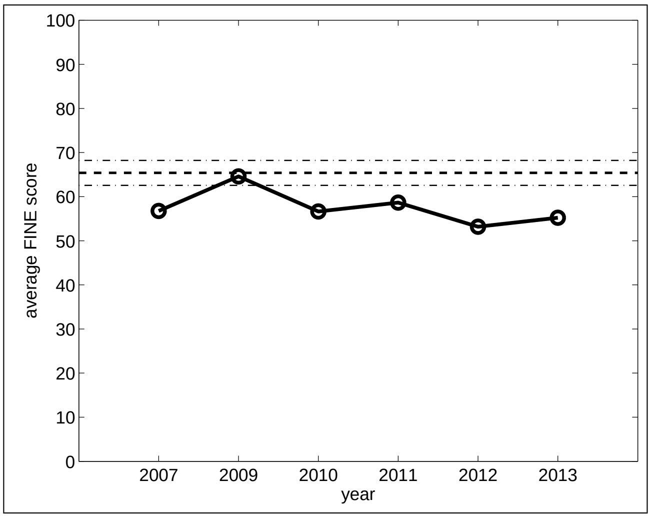

Figure 2. Average FINE score of best performing system (y-axis) vs. year (x-axis) plotted as solid line. Upper bound plus confidence interval plotted as dashed line.

year	system	mean	var	t
2007	PS	56.75	848.09	-4.3475
2009	PS2	64.58	633.76	-0.4415
2010	SSPK2	56.64	726.78	-4.6230
2011	SSPK2	58.64	687.91	-3.6248
2012	SSKS2	53.19	783.44	-6.3018
2013	SS2	55.21	692.23	-5.4604

Table 3. Comparison of best system vs. upper bound due to lack of inter-rater agreement.

Interestingly, this system PS2 which gave the peak performance of all AMS years has also participated in 2010 to 2013. In terms of statistical significance (as measured via Friedman tests as part of the MIREX evaluation), PS2 has performed on the same level with the top systems of all following years. The systems PS2 has been submitted by Tim Pohle and Dominik Schnitzer and essentially consists of a timbre and a rhythm component [12]. Its main ingredients are MFCCs modeled via single Gaussians and Fluctuation patterns. It also uses the so-called P-norm normalization of distance spaces for combination of timbre and rhythm and to reduce the effect of hubness (anormal behavior of distance spaces due to high dimensionality, see [6] for a discussion related to the AMS task and [14] on re-scaling of distance spaces to avoid these effects).

As outlined in Section 3, from 2007 on the same database of songs was used for the AMS tasks. However, each year a different set of 100 or 50 songs was chosen for the human listening tests. This fact can explain that the one algorithm participating from 2009 to 2013 did not always perform at the exact same level. After all, not only the choice of different human graders is a source of variance in the obtained FINE scores, but also the choice of different song material. However, the fact that the one algorithm that reached the upper bound has so far not been outperformed adds additional evidence that the upper bound that we obtained indeed is valid.

5. DISCUSSION

Our meta analysis of all editions of the MIREX "Audio Music Similarity and Retrieval" tasks conducted so far has produced somewhat sobering results. Due to the lack of inter-rater agreement there exists an upper bound of performance in subjective evaluation of music similarity. Such an upper bound will always exist when a number of different people have to agree on a concept as complex as that of music similarity. The fact that in the MIREX AMS task the notion of similarity is not defined very clearly adds to this general problem. After all, to "sound similar" does mean something quite different to different people listening to diverse music. As a consequence, an algorithm that has reached this upper bound of performance already in 2009 has not been outperformed ever since. Following our argumentation, this algorithm cannot be outperformed since any additional performance will be lost in the variance of the different human graders.

We now like to discuss a number of recommendations for future editions of the AMS task. One possibility is to go back to the procedure of AMS 2006 and again have more than one grader rate the same query/candidate pairs. This would allow to always also quantify the degree of inter-rater agreement and obtain upper bounds specific to the respective test songs. As we have argued above, we believe that the upper bound we obtained for AMS 2006 is valid for all AMS tasks. Therefore obtaining specific upper bounds would make much more sense if future AMS tasks would use an entirely different database of music. Such a change of song material would be a healthy choice in any case. Re-introducing multiple ratings per query/candidate pair would of course multiply the work load and effort if the number of song pairs to be evaluated should stay the same. However, using so-called "minimal test collections"-algorithms allows to obtain accurate estimates on much reduced numbers of query/candidate pairs as has already been demonstrated for the AMS task [18]. In addition rater-specific normalization should be explored. While some human graders use the full range of available FINE scores when grading similarity of song pairs, others might e.g. never rate song pairs as being very similar or not similar at all, thereby staying away from the extremes of the scale. Such differences in rating style could add even more variance to the overall task and should therefore be taken care of via normalization.

However, all this would still not change the fundamental problem that the concept of music similarity is formulated in such a diffuse way that high inter-rater agreement cannot be expected. Therefore, it is probably necessary to research what the concept of music similarity actually means to human listeners. Such an exploration of what perceptual qualities are relevant to human listeners has already been conducted in the MIR community for the specific case of textural sounds [7]. Textural sounds are sounds that appear stationary as opposed to evolving over time and are therefore much simpler and constrained than real songs. By conducting mixed qualitative-quantitative interviews the authors were able to show that qualities like "high-low", "smooth-coarse" or "tonal-noisy" are important to humans discerning textural sounds. A similar approach could be explored for real song material, probably starting with a limited subset of genres. After such perceptual qualities have then been identified, future AMS tasks could ask human graders how similar pairs of songs are according to a specific quality of the music. Such qualities might not necessarily be straight forward musical concepts like melody, rhythm, or tempo, but rather more abstract notions like instrumentation, genre or specific recording effects signifying a certain style. Such a more fine-grained approach to music similarity would hopefully raise inter-rater agreement and make more room for improvements in modeling music similarity.

Last but not least it has been noted repeatedly that evaluation of abstract music similarity detached from a specific user scenario and corresponding user needs might not be meaningful at all [13]. Instead the MIR community might have to change to evaluation of complete music retrieval systems, thereby opening a whole new chapter for MIR research. Such an evaluation of a complete real life MIR system could center around a specific task for the users (e.g. building a playlist or finding specific music) thereby making the goal of the evaluation much clearer. Incidentally, this has already been named as one of the grand challenges for future MIR research [15]. And even more importantly, exactly such a user centered evaluation will happen at this year's tenth MIREX anniversary: the "MIREX Grand Challenge 2014: User Experience (GC14UX)" [4]. The task for participating teams is to create a web-based interface that supports users looking for background music for a short video. Systems will be rated by human evaluators on a number of important criteria with respect to user experience.

6. CONCLUSION

In our paper we have raised the important issue of the lack of inter-rater agreement in human evaluation of music information retrieval systems. Since human appraisal of phenomena as complex and multi-dimensional as music sim-

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

ilarity is highly subjective and depends on many factors such as personal preferences and past experiences, evaluation based on human judgments naturally shows high variance across subjects. This lack of inter-rater agreement presents a natural upper bound for performance of automatic analysis systems. We have demonstrated and analysed this problem in the context of the MIREX ”Audio Music Similarity and Retrieval” task, but any evaluation of MIR systems that is based on ground truth annotated by humans has the same fundamental problem. Other examples from the MIREX campaign include such diverse tasks as ”Structural Segmentation”, ”Symbolic Melodic Similarity” or ”Audio Classification”, which are all based on human annotations of varying degrees of ambiguity. Future research should explore upper bounds of performance for these many other MIR tasks based on human annotated data.

7. ACKNOWLEDGEMENTS

We would like to thank all the spiffy people who have made the MIREX evaluation campaign possible over the last ten years, including of course J. Stephen Downie and his people at IMIRSEL. This work was supported by the Austrian Science Fund (FWF, grants P27082 and Z159).

8. REFERENCES

[1] Cohen J.: Statistical power analysis for the behavioral sciences, L. Erlbaum Associates, Second Edition, 1988.

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

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

[4] Fleiss J.L.: Measuring nominal scale agreement among many raters, Psychological Bulletin, Vol. 76(5), pp. 378-382, 1971.

[5] Flexer A., Schnitzer D.: Effects of Album and Artist Filters in Audio Similarity Computed for Very Large Music Databases, Computer Music Journal, Volume 34, Number 3, pp. 20-28, 2010.

[6] Flexer A., Schnitzer D., Schlüter J.: A MIREX meta-analysis of hubness in audio music similarity, Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR’12), 2012.

[7] Grill T., Flexer A., Cunningham S.: Identification of perceptual qualities in textural sounds using the repertory grid method, in Proceedings of the 6th Audio Mostly Conference, Coimbra, Portugal, 2011.

[8] Jones M.C., Downie J.S., Ehmann A.F.: Human Similarity Judgments: Implications for the Design of Formal Evaluations, in Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR’07), pp. 539-542, 2007.

[9] Landis J.R., Koch G.G.: The measurement of observer agreement for categorical data, Biometrics, Vol. 33, pp. 159174, 1977.

[10] Novello A., McKinney M.F., Kohlrausch A.: Perceptual Evaluation of Music Similarity, Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR 2006), Victoria, Canada, 2006.

[11] Pampalk E.: Computational Models of Music Similarity and their Application to Music Information Retrieval, Vienna University of Technology, Austria, Doctoral Thesis, 2006.

[12] Pohle T., Schnitzer D., Schedl M., Knees P., Widmer G.: On Rhythm and General Music Similarity, Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR09), 2009.

[13] 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.

[14] Schnitzer D., Flexer A., Schedl M., Widmer G.: Local and Global Scaling Reduce Hubs in Space, Journal of Machine Learning Research, 13(Oct):2871-2902, 2012.

[15] Serra 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. (editor), 2013.

[16] Sturm B.L.: Classification accuracy is not enough, Journal of Intelligent Information Systems, 41(3), pp. 371-406, 2013.

[17] Urbano J., Downie J.S., McFee B., Schedl M.: How Significant is Statistically Significant? The case of Audio Music Similarity and Retrieval, in Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR’12), pp. 181-186, 2012.

[18] Urbano J., Schedl M.: Minimal test collections for low-cost evaluation of audio music similarity and retrieval systems, International Journal of Multimedia Information Retrieval, 2(1), pp. 59-70, 2013.

[19] Vignoli F.: Digital Music Interaction Concepts: A User Study, Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR’04), Barcelona, Spain, 2004.

[20] West K.: Novel techniques for audio music classification and search, PhD thesis, University of East Anglia, 2008.