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.