HUMAN AND AUTOMATED JUDGEMENTS OF SIMILARITY IN A GLOBAL MUSIC SAMPLE: A PRELIMINARY ANALYSIS

Ding Shenghao[1] **Hideo Daikoku**[1] **Ujwal Sriharsha Sanne**[2]
Marino Kinoshita[1] **Rei Konno**[1] **Yoichi Kitayama**[1]
Shinya Fujii[1] **Patrick E. Savage**[1]
[1] Keio University, Japan
[2] Centre for Digital Music, Queen Mary University of London, UK
psavage@sfc.keio.ac.jp

EXTENDED ABSTRACT

Recent research in Music Information Retrieval (MIR) makes automatic of musical similarity possible [1]. Expanding such automated analyses across cultures, however, is still a very challenging topic [8]. Several researchers have annotated performance-style characteristics of thousands of recordings of traditional music from around the world using manual classification schemes such as 'Cantometrics', but because they are classified manually the ratings are subjective and time-consuming [2] [3] [7] [6] [4]. Panteli et al. proposed an automated algorithm [5] that could overcome these limitations and applied it to in automatic analysis on world music similarities, but its validity is unclear because they were not able to test it against ground-truth human annotations. This paper attempts to combine both human ground-truth ratings and automated algorithms to evaluate the reliability of human ratings and evaluate Panteli et al.'s automated algorithm against ground-truth data.

The musical sample used in the study is the 30 song 'Consensus Tape' subset used as the test set for Lomax's Cantometrics Training Tapes [3]. We implemented three methods to retrieve the similarity information from 62 examinees and divided all 30 songs into six groups of 5 songs.

After the examinee gets a brief training using examples from the Cantometrics Training Tape, they answer questions about the 5 songs using three methods: 1) Song-wise: Each song is rated on a 3-point Likert scale for 6 variables - Ornamentation, Vocal Range, Tempo, Rhythmic Regularity, Vocal Tension, and Vocal Texture - modified slightly from Cantometrics [3]. 2) Pair-wise: Each of 10 possible pairs of 5 songs are rated for their overall similarity on a 5-point Likert scale. 3) Triplet: Each of the 10 possible combinations of 3 songs from within the 5 are rated in an 'odd-one-out' paradigm by asking the listener to choose the song that is most different from the others. Song order was randomized within each of the three blocks, and the order of pairwise and triplet blocks was randomized (song-wise comparisons were always first). We retrieved the data from 64 people, and Figure 1 shows the average partial distance matrices for each of the three methods as well as the distance matrix based on Panteli et al.'s algorithm.

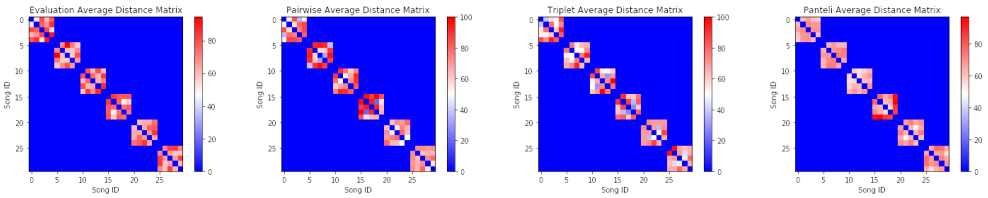

Figure 1. Average distance matrices among the 30 songs across different methods

Figure 2 shows that people may agree with each other when analyzing the similarities with the same methods. However, the correlation across methods is unstable. While the correlation between Featurewise Evaluation and Pairwise comparison are strong in Group 1 (r = 0.76) and Group 3 (r = 0.85), most of the other groups and methods have low correlations. In future analyses we plan to more thoroughly investigate reliability within and between groups of human annotators and automated algorithms, but we suspect that agreement between perceptual ratings and MIR algorithms may remain weak until we develop new algorithms and features based on more cross-culturally applicable musical theories.

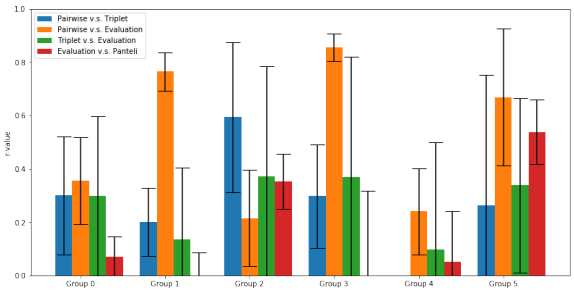

Figure 2. Correlations between different distance matrices for different groups

ACKNOWLEDGMENTS

We thank Ujwal Sriharsha Sanne for extensive help implementing the automated similarity analyses. Ujwal tragically passed away only a few weeks before submission, so he was unable to read and approve the submission, but we have included him as an author in order to acknowledge his work on the project and respect his memory. We thank the YAMAHA corporation (particularly Taiki Shimozono, Rento Tanase, Satoshi Usa, and Takuya Fujishima) for providing support, feedback, and experiment participants. Additional funding for this study was provided by Startup Grants from the Keio Global Research Institute and Keio Research Institute at SFC awarded to P.E.S.

REFERENCES

[1] Michael Casey, Remco Veltkamp, Masataka Goto, Marc Leman, Christophe Rhodes, and Malcolm Slaney. Content-based music information retrieval: Current directions and future challenges. *Proceedings of the IEEE*, 96:668 – 696, 05 2008.

[2] Alan Lomax. *Folk song style and culture*. American Association for the Advancement of Science, 1968.

[3] Alan Lomax. *Cantometrics: An Approach to the Anthropology of Music*. Berkeley: Uneversity of California Extension Media Center, 1976.

[4] Samuel A Mehr, Manvir Singh, Dean Knox, Daniel Ketter, Daniel Pickens-Jones, S. Atwood, Christopher Lucas, Nori Jacoby, Alena Egner, Erin J Hopkins, and et al. Universality and diversity in human song, Nov 2018.

[5] Maria Panteli, Emmanouil Benetos, and Simon Dixon. A computational study on outliers in world music. *PLOS ONE*, 12(12):e0189399, December 2017.

[6] Patrick E. Savage. Alan lomaxs cantometrics project: A comprehensive review. *Music & Science*, 1:2059204318786084, 2018.

[7] Patrick E. Savage, Steven Brown, Emi Sakai, and Thomas E. Currie. Statistical universals reveal the structures and functions of human music. *Proceedings of the National Academy of Sciences*, 112(29):8987–8992, 2015.

[8] George Tzanetakis, W. Schloss, and Mathew Wright. Computational ethnomusicology. *Journal of Interdisciplinary Music Studies*, 1:1–24, 01 2007.