AGREEMENT AMONG HUMAN AND AUTOMATED ESTIMATES OF SIMILARITY IN A GLOBAL MUSIC SAMPLE

[1]Hideo Daikoku, [1]Shenghao Ding, [2]Emmanouil Benetos, [3]Anna Lomax Chairetakis Wood
[4]Taiki Shimizono,[2]Ujwal Sriharsha Sanne, [1]Shinya Fujii, [1]Patrick E. Savage*
[1]Keio University, Japan, [2]Queen Mary University London,
[3]Association for Cultural Equity, [4]Yamaha Corporation, Japan
*psavage@sfc.keio.ac.jp

ABSTRACT

While music information retrieval (MIR) has made substantial progress in automatic analysis of audio similarity for Western music, it remains unclear whether these algorithms can be meaningfully applied to cross-cultural analyses of more diverse musics. Here we collect perceptual ratings from 62 Japanese participants using a global sample of 30 traditional songs, and compare these ratings against both pre-existing expert annotations and audio similarity algorithms. We find that different methods of perceptual ratings all produced similar, moderate levels of inter-rater agreement comparable to previous studies, but that agreement between human and automated methods is always low regardless of the specific methods used to calculate musical similarity. Our findings suggest that the MIR methods tested are unable to measure cross-cultural music similarity in perceptually meaningful ways.

1. INTRODUCTION

Advances in MIR have paved the way for rapid and relatively accurate automatic analysis of Western classical and popular music (Casey et al., 2008). MIR has achieved this success in part through developing algorithms that incorporate aspects of Western music theory such as 12-tone equal tempered scales and 4/4 meter. In contrast, there have been few automatic studies of non-Western music, for which such theories do not necessarily apply (Panteli et al., 2017, 2018; Tzanetakis et al., 2007). Crucially, no MIR studies have successfully validated automatic analysis of a culturally diverse musical dataset against human perceptual data. Thus, it remains unknown whether current MIR technologies can meaningfully be used to automatically compare diverse music from throughout the world.

Ethnomusicologists have spent over a century developing manual methods of comparing music from around the world by manually annotating recordings, evaluating features by ear and classifying them subjectively (Ellis, 1885; Nettl & Bohlman, 1991; Savage & Brown, 2013). While various manual systems of cross-cultural classification have been developed, they all have their own drawbacks (Lomax, 1968; Savage, 2018; Tenzer & Roeder, 2011; Savage & Brown, 2013; Panteli et al., 2017). In particular, manual evaluation is subjective and time-consuming, two problems that could potentially be overcome through successfully automating the evaluation process.

The notion of musical similarity is a central issue spanning MIR, music cognition, and musicology. Because concepts of musical similarity vary both within and between cultures, there is no single objective measure of musical similarity that can be used to evaluate similarity algorithms (Allan et al., 2007). Indeed, limited inter-rater agreement has been proposed to represent an upper bound limiting progress in MIR algorithms (Flexer & Grill, 2016). Inter-rater agreement has also been raised as an issue potentially limiting cross-cultural musicological analyses (Savage, 2018; Mehr et al., 2019). Evaluating automated algorithms against human ratings of cross-cultural similarity thus requires collecting perceptual data from multiple raters and evaluating their inter-rater agreement.

The main goal of our paper is to evaluate the ability of automated algorithms to match perceptual similarity for a global dataset of music. We collect perceptual data from 62 participants using three different methods of similarity ratings for 30 diverse traditional songs from around the world, and compare these with previously published manual annotations and with two publicly available automated audio similarity algorithms. Overall, we find moderate agreement among human raters but no agreement between human and existing automated algorithms, suggesting the need for developing cross-culturally valid automated methods in future research.

Section 2 describes the music dataset used in this study. Section 3 analyzes the inter-rater agreement of our novel perceptual data and compares it with previous studies. Section 4 compares similarity measurements from our human and automated measurements of musical similarity. Section 6 discusses these results and their limitations, and proposes future work toward collecting data for cross-validation and creating new computational methods for analyzing all the world's music.

1.1 Related work

1.1.1 Human judgments of musical similarity

A landmark attempt to measure cross-cultural musical similarity was Alan Lomax's Cantometrics Project (Canto = Song, Metrics = Measure) (Lomax, 1968; Savage, 2018; Wood, 2018; Wood et al., 2021). Lomax and his colleagues analysed thousands of songs from hundreds of worldwide

societies using 37 classificatory features, and compared the resulting patterns of similarities and differences with aspects of social structure and cultural history. These diverse features span domains such as song structure (e.g., melodic range), singing style (e.g., vocal width), and social context (e.g., solo/group arrangement of singers). The recent digitization and publication of over 5,000 Cantometrics codings and accompanying audio has made it possible to apply it to larger-scale automatic algorithms in the context of music similarity (Wood et al., 2021). Recent analysis of the original Cantometrics codings has shown promise in using it to explore musical diversity in India (Daikoku et al., 2020). Here we choose Cantometrics for its ability to capture variation cross-culturally in diverse musical styles in a large pre-annotated dataset.

Constructive criticism of aspects of Cantometrics such as inter-rater agreement and calculation of musical similarity led to attempts to design more reliable methods of classifying and comparing music cross-culturally (Savage et al., 2012; Rzeszutek et al., 2012; Savage, 2018; Mehr et al., 2019; Proutskova, 2019). CantoCore, a classification scheme inspired by Cantometrics, was found to have higher average inter-rater agreement (measured by Cohen's Kappa) than Cantometrics when compared against a dataset of 30 traditional songs from around the world (Savage et al., 2012). Mehr et al. (Mehr et al., 2019) argued that the reliability of their own scheme (measured by Cronbach's Alpha) was in turn higher than both Cantometrics and CantoCore [1]. Independent of the reliability of these schemes, the methods for converting these classification schemes into measurements of overall musical similarity have yet to be validated against perceptual ratings of musical similarity.

1.1.2 Automatic audio similarity

Audio signal processing and machine learning have made it possible to use computational tools to quantify musical similarity directly from audio files using features such as Mel Frequency Cepstral Coefficients (MFCCs) (Urbano et al., 2014), tempograms (Grosche & Müller, 2011), pitch bi-histograms (Van Balen et al., 2014) and chromagrams (Pauws, 2004) for timbre, rhythm, melody and harmony respectively.

The MIREX campaign had a category specifically for Audio Similarity and Retrieval [2] from 2006-2014 [3], during which time several research groups made progress in accuracy for the datasets provided. One algorithm

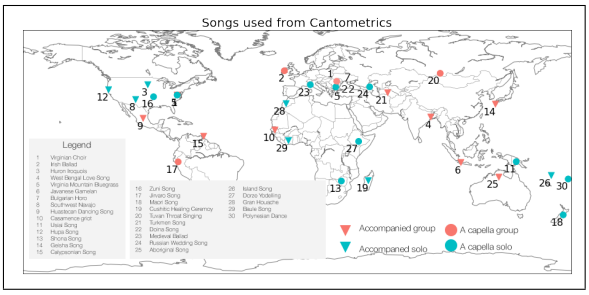

Figure 1: A map showing the approximate locations and cultural group names of the 30-song sample from the Cantometrics Consensus Tape (adapted from (Lomax, 1976; Savage et al., 2012)). We label songs by solo vs. group singing and with vs. without instrumental accompaniment.

that consistently performed at or near the top in multiple MIREX evaluation years was an algorithm that emphasizes aspects of timbre and rhythm modelled by MFCCs, and has been implemented in the commercial software Musly (Pohle et al., 2009; Schnitzer et al., 2011; Flexer & Grill, 2016). However its accuracy for non-Western music has yet to be systematically evaluated.

More recently, Panteli et al. proposed an automated algorithm (Panteli et al., 2017) specifically in order to measure similarity in global samples including non-Western music. To do so, they extracted features related to rhythm (onset patterns with scale transform), melody (pitch bi-histograms), harmony (average chromagrams, using "20-cent pitch resolution to allow for microtonality"), and timbre (MFCCs), analyzed with linear discriminant analysis (LDA) to identify musical outliers (i.e., songs that were very musically dissimilar). Because Panteli et al. lacked ground-truth perceptual data, they relied on country-of-origin labels to train and test the accuracy of its musical similarity algorithm. However, this assumes that the amount of musical diversity within a country is fairly small relative to differences between cultures, whereas analyses of cross-cultural musical diversity have instead found that within-culture variation tends to be greater than between-culture variation (Rzeszutek et al., 2012; Mehr et al., 2019; Daikoku et al., 2020). Thus there is still a need to evaluate Panteli et al.'s algorithm against human perceptual data on musical similarity.

2. CROSS-CULTURAL MUSIC DATASET

We chose the 30 audio recordings of traditional songs from Lomax's Cantometrics Consensus Tape (Lomax, 1968) as the global dataset for our study. Here we use the word 'global dataset' to refer to the cross-cultural music dataset chosen, and refer to each 'song' as the randomly selected excerpts of these recordings. These recordings were originally used by Lomax to test human raters after training them to use the 37 Cantometrics features (such as vocal texture, melodic range, vocal tension, tempo, rhythmic regularity, and ornamentation) to classify songs around the world, and thus already functioned as one type of pre-annotated expert ground-truth data (Lomax, 1976; Savage

[1] This comparison is problematic because not only can Alpha statistics not be directly compared with Kappa statistics, but Cronbach's Alpha is a function of the number of raters, so collecting data from large numbers of raters (e.g., 30 raters in (Mehr et al., 2019)) gives an inflated appearance of inter-rater agreement compared to smaller numbers of raters (e.g., 2 raters in (Savage et al., 2012, 2015)) even if the level of agreement between each pair of raters is the same (Revelle & Condon, 2019).

[2] MIREX also evaluates symbolic music similarity as represented through staff notation. However, the reliability of such notation for non-Western music is debated (List, 1974) and has not yet been objectively evaluated, so we have not included symbolic notation in the current study.

[3] Flexer and Grill (Flexer & Grill, 2016) explain that this section was discontinued from MIREX from 2015 because "only our own research team, again sending the same peak performing system PS2 (Pohle et al., 2009) for the seventh year, wanted to participate".

et al., 2012). Not only were these 30 songs pre-annotated using Cantometrics, but almost 6,000 more songs have also been annotated using Cantometrics (Lomax, 1968; Savage, 2018; Wood, 2018; Wood et al., 2021), making Cantometrics ratings a valuable ground-truth data set for comparison. We focus on only 30 songs and only use 6 of the 37 original Cantometric features due to limitations in experiment length and listener fatigue (motivations for specific feature choices are explained in Section 3).

These recordings were chosen for their uniqueness in style and representativeness of a broad variety of musical cultures. Each song is from a different cultural group and has some shared and some distinct characteristics with other songs in the sample. This sample was also previously used to compare reliability in Cantometric ratings against the CantoCore rating scheme, which was inspired by Cantometrics but focused on aspects of song structure rather than singing style (Savage et al., 2012). Because the structure of the singing group and instrumental accompaniment (solo vs. group singing, a cappella vs. instrumental accompaniment) were previously found to be important markers of overall style, the recordings in Figure 1 and the other figures have been labeled to highlight these contrasts (Savage et al., 2015; Daikoku et al., 2020). The original recordings were excerpts of between 40 seconds to 2 minutes 20 seconds, but to make our experiments feasible and enable participants to remember the sound of multiple recordings in order to compare them, we randomly selected short 10-second excerpts containing singing from each recording. The same 10-second excerpts were used for all experiments.

2.1 Participants

We tested 62 participants employed at Yamaha's headquarters in Japan. Participant ages ranged from 25-63, out of which 90% spoke Japanese as their native language, and 81% understood English as either a primary or secondary language. The test was conducted fully in Japanese with the option to change languages to English or Chinese. 62% of all participants had played a musical instrument for over 10 years, while 16% had no musical training.

3. PERCEPTUAL EXPERIMENT

3.1 Experiment design

Collecting a large enough sample of perceptual ratings to be able to evaluate both inter-rater agreement and human-automated reliability is challenging. For example, for listeners to rate 5 recordings requires making judgments of sets of individual features for 5 recordings (1, 2, 3, 4, 5), similarity for 10 different pairs (1 vs. 2, 1 vs. 3, 2 vs. 3, etc.) or 10 different triplets (1 vs. 2 vs. 3, 1 vs. 2 vs. 4, etc.), which takes approximately 30 minutes. However, human judgments for only 5 recordings would not be enough to meaningfully compare with automated algorithms. On the other hand, increasing the sample to 10 recordings would require rating 10 sets of features, 45 pairs, and 120 triplets, which is already more than can be collected within the course of a 1-hour experiment, especially when accounting for listener fatigue. If we attempt to spread out the data collection across multiple different participants by having different participants rate different recordings, we lose the ability to compare inter-rater agreement between participants. Unfamiliarity, use/absence of reference tracks, and order effects can also affect perception of similarity.

To balance the need for enough data to compare both human-human and human-automated agreement, we designed an experiment where we divided the set of 30 diverse recordings previously used to evaluate inter-rater agreement into 6 sets of 5 recordings. For each set, we collected perceptual judgments of all possible features, pairs, and triplets from 10-11 participants per set (total $n = 62$ participants). The 62 participants were divided into 6 groups, where all members within each group rated the same 5 songs from the 30-song dataset.

Each experiment lasted approximately 20-30 minutes and was divided into three blocks: feature evaluation, pairwise evaluation and triplet (odd-one-out) evaluation. Before beginning the experiment, participants are played a series of reference tracks taken from the Cantometrics training tapes in order to familiarize them with the features they would be rating and the types of recordings they would be asked to rate. Participants then evaluate a set of features for each song after listening to each song at least once, after which they performed the triplet and pairwise similarity tasks. The order of the triplet and pairwise blocks, and the order of songs/combinations within each block was randomized so as to negate order effects, but the feature evaluation block always came before the triplet and pairwise blocks in order to familiarize participants with the set of 5 recordings before asking them to rate similarity among recordings. Although the experiment interface was accessed online, all participants were monitored in person to maximize data quality. After the experiment was over, participants were asked to fill out a questionnaire asking them about their age, gender, musical experience, preferences, and exposure to non-Western music. Please refer to the supplementary material for details on experiment design.

4. HUMAN VS. AUTOMATED JUDGMENTS OF MUSICAL SIMILARITY

Having determined inter-rater agreement for our human judgments of similarity, we proceed to compare human judgments vs. automatic audio similarity algorithms. To do so, we created the following five distance matrices and compared them against one another: H1a: Expert (Cantometric nominal); H1b: Expert (Cantometric ordinal); H2: Naive listener (5x5 pairwise ratings x 6); A1: Musly (Schnitzer et al., 2011); A2: Panteli et al. (Panteli et al., 2017) ("H" indicates human ratings, "A" indicates automated judgments). All matrices are 30x30 distance matrices, except for "H2", which is only a partial distance matrix containing the six 5x5 matrices of pairwise distances collected from the perceptual experiment

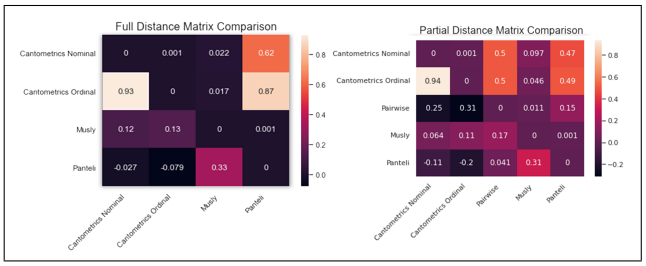

Figure 2: Distance matrix correlations among different methods of measuring musical similarity for the full 30x30 distance matrix and the partial distance matrix containing six 5x5 distance matrices. The bottom left triangle indicates the correlation coefficient, and the top right triangle shows its corresponding p-value. Ordinal Cantometrics features are ordered features like tremolo where values range from little to much tremolo; Nominal Cantometrics features are unordered like vocal organization where values can describe solo, unison, heterophony or polyphony.

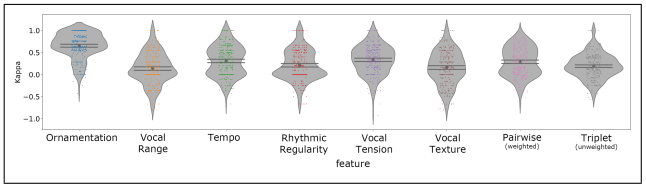

Figure 3: Violin plots showing inter-rater reliability (Kappa) for the different rating methods used in the current study: pairwise, triplet, and feature wise for six different stylistic features. Each individual point ($n = 290-300$ per condition) represents a Kappa value for a given pair of participants. Large dots represent means, horizontal lines represent 95% confidence intervals. See supplementary materials for more.

(because it was not practical to collect perceptual ratings of pairwise similarity for all 435 pairs among the 30 songs for each participant).

4.1 Distance matrix correlations

Because H2 could not include the full 30x30 distance matrix, we performed two sets of distance matrix correlations (Figure 2). The first set compared only the six 5x5 matrices containing all pairwise distances within each of the six groups of 5 songs, using pairwise similarity (H2) from our perceptual experiments and excluding between-group data from the other distance matrices (H1a, H1b, A1, and A2). The second set compared full 30x30 distance matrices, excluding the partial matrix H2.

Both sets of comparisons used Mantel's permutation-based test of distance matrix correlations to control for non-independence of data points within distance matrices (Legendre & Legendre, 2012). The Mantel test involves repeatedly testing the correlations with random permutations of the rows and columns of one of the matrices. Statistical significance is calculated from the proportion of permutations that lead to a higher correlation coefficient (Mantel, 1967).

The distance matrix correlations suggest minimal correlation between human ground-truth ratings of similarity and automated algorithmic calculations of similarity. While the two automated methods are moderately correlated with one another (r=.33, p<.001) and the three human annotated methods are moderately or strongly correlated with one another (H1a-H1b: r=.93, p<.001; H1a-H2: r=.25, p=.02; H1b-H2: r=.31, p=.005), all of the correlations between the automated distance matrices and human ground-truth distance matrices are weak (r <.2).

4.2 Human-human agreement

We calculated Kappa statistics (Cohen, 1968), a metric often used to assess the agreement between two raters, to determine inter-rater agreement of our perceptual data. Here 0 indicates agreement equivalent to chance, 0.41 – 0.60 indicates moderate agreement, and 1 indicates perfect agreement. Quadratic weighted Kappa is calculated for feature-wise and pairwise ratings, because these are ordinal Likert scales, while unweighted Kappa is calculated for triplet ratings, because these are nominal (unordered) ratings. Kappa was originally developed for only two raters, but can be extended by calculating all possible pairwise Kappa values for larger numbers of raters (Light, 1971). Our 62 participants were distributed into 6 groups of 10-11 participants, and within each group all participants rated the same combinations of 5 songs. This resulted in a total of 45-55 pairwise Kappa values per group, for a total of 290-300 pairwise Kappa values for each condition (See figure).

5. LIMITATIONS

While our global music sample is highly diverse, our sample of 62 participants were recruited solely from one company because the company offered to support our project, including participant recruitment. Therefore, while our participant sample avoids the traditional bias in music cognition studies (Jacoby et al., 2020) towards homogeneous "WEIRD" (Western, Educated, Industrialized, Rich, Democratic (Henrich et al., 2010)) samples of Western undergraduate students, it does not necessarily generalize to the broader Japanese population let alone to all humans. We have begun to expand this paradigm to include more diverse participants from different societies around the world to more systematically investigate the degree to which musical perception and cognition vary within and between cultures (Jacoby & McDermott, 2017; Jacoby et al., 2020).

We briefly trained all participants and had them rate songs using 6 Cantometric features prior to collecting pairwise and triplet similarity ratings, to ensure that participants had listened to all relevant songs prior to their similarity ratings to maximize reliability of ratings. However, it is possible that this primed participants to influence their judgments of similarity to more strongly weight these 6 features. Such influence seems unlikely to fully explain why similarity ratings based on all 37 Cantometric features for longer 1-2-minute song excerpts were more strongly

correlated with these ratings (primed with only 6 features and tested on only 10-second randomly selected clips) than automated ratings were. However, future experiments exploring the effects of reversing or fully randomize this order may clarify how such priming might affect results (although it should be noted that this may conceivably reduce inter-rater reliability to levels that may render similarity judgments essentially arbitrary). Techniques such as metric learning (McFee et al., 2012; Wolff & Weyde, 2014) may also be useful in future research to investigate how these 6 features and/or other individual features are cognitively weighted to give overall similarity judgments (our current method of calculating musical similarity weighted all individual features equally (Rzeszutek et al., 2012)).

The two audio similarity algorithms evaluated in this study were chosen because they either performed best in MIREX's audio similarity evaluation, or in (Panteli et al., 2017)'s case were designed specifically for measuring similarity in non-Western music. However, these systems have limitations, such as emphasis on timbral features (Schnitzer et al., 2011) and reliance on country labels as a proxy for similarity (Panteli et al., 2017) that may contribute to their poor performance. In the future we would like to explore alternative methods such as existing Gaussian mixture models (Jensen et al., 2007), metric learning (McFee et al., 2012, 2011), and deep learning (Cheng et al., 2020). We would also like to develop new models explicitly based on cross-culturally motivated features and data, and such attempts may now be feasible with the recent publication of the Cantometrics dataset of over 5,000 expert-annotated audio recordings (Wood et al., 2021).

6. DISCUSSION

We found no major differences in inter-rater reliability between perceptual ratings based on feature-wise vs. pairwise comparisons. This suggests that data from either of these types of ratings is likely to be of value in future large-scale cross-cultural analyses. This means it may be feasible to analyze large databases such as Cantometrics, which contain feature wise expert ratings for over 5,000 audio recordings (Lomax, 1968; Savage, 2018; Wood, 2018; Wood et al., 2021). Having validated Cantometric ratings against human perceptual data of musical similarity opens up the possibility of using larger datasets of full Cantometrics data as training data for future supervised learning approaches (Cheng et al., 2020; McFee et al., 2012; Jensen et al., 2007). As machine learning approaches improve, such supervised learning may offer better chances of matching human judgments than the feature-extraction approach used in the current study.

This paper presents a new dataset for similarity data in a global music sample. We provide analysis of agreement between the data collected from experts, non-experts, and feature-based estimations of similarity. We also compare three methods for collecting ground truth data and hope it will be useful for data collection and inter-rater agreement. Our results show low levels of agreement in measurements of similarity between human and automated judgments of music similarity in a global musical sample. This is consistent with previous arguments that current MIR methods are not yet suitable for global samples including non-Western music (Tzanetakis et al., 2007). This is also consistent with previous studies that have suggested that automated algorithms are fundamentally limited in their ability to measure musical similarity due to the subjective nature of human perceptions of similarity even for Western music (Flexer & Grill, 2016). However, our finding of moderate levels of inter-rater agreement among human raters comparable with previous studies suggests that limited inter-rater agreement cannot be the sole reason for the poor performance of the automated algorithms.

Indeed, pairwise inter-rater agreement values for the current study (mean Kappa = .29) were similar to those previously reported for Western pop music (Kappa = .21-.29) (Jones et al., 2007). These levels are significantly above chance, but not particularly high (some have suggested .21-0.4 is 'fair' (Landis & Koch, 1977), while others argue less than 0.4 is unacceptable for certain applications (e.g., clinical diagnoses) (Sim & Wright, 2005). Due to logistical constraints we only focused on collecting perceptual data for a global music sample, however, in order to evaluate the relative reliability of MIR methods on Western vs. non-Western music, we would need a controlled comparison using samples of Western music as well as non-Western music.

Notably, a recent cross-cultural study of popular music in three countries (USA, Brazil, and Korea) found substantial agreement (r=0.49-0.63) between human and automated estimates of mood (e.g., "danceability", "sadness") (Lee et al., 2021). This suggests that cross-culturally meaningful MIR algorithms are not inherently impossible. Future study will be needed to determine whether the lower reliability we found is due to "similarity" being a more multidimensional concept than mood concepts like "danceability" and "sadness", due to the more diverse sample of traditional music in our corpus, or due to other differences in experimental design.

Inter-rater agreement in our current study of amateur participants with a diverse global sample of traditional music was comparable to previous studies using expert musicologists and/or Western popular music (Savage et al., 2012). This suggests that there is nothing inherently insurmountable about cross-cultural comparison of music (Savage & Brown, 2013), since even amateurs without any experience listening to diverse music from around the world were able to give reasonably reliable ratings after a brief ($\sim$5-minute) training period at the beginning of the experiment. This suggests that future cross-cultural studies do not necessarily need to rely only on annotations by expert musicologists, opening possibilities for larger-scale studies (e.g., crowd-sourced online experiments). Crucially, this study provides insight into the framework necessary for collecting ground truth data while minimizing experiment time and the total number of comparisons each participant would need to do to successfully evaluate similarity. Future work could build upon this framework for larger

studies and a more diverse participant pool.

If we can succeed in developing and validating methods for automated analysis of all the world's music, it could open up new ways to help people find and appreciate diverse music throughout the world. We hope to apply such findings to fields such as information science, music cognition, anthropology, and to broader applications including music recommendation services, music copyright law, music education, cultural heritage preservation, and cross-cultural understanding through music.

7. AUTHOR CONTRIBUTIONS

Conceptualization: PES, SF, EB, HD, ALCW, TS; Methodology/Analysis/Investigation/Visualization; HD, DS, USS, PES; Project administration/supervision/funding acquisition: PES; Writing - original draft: HD, PES. Writing - review & editing: EB, SF, ALCW

8. ACKNOWLEDGMENTS

Ujwal Sriharsha Sanne tragically passed away in August 2019, 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 M. Kinoshita and R. Konno for assistance during early phases of experiment design; and Polina Proutskova, Victor Grauer, Simon Dixon, Marcus Pearce, Peter Harrison, and Adrien Ycart for comments on earlier versions of this manuscript. This is an expanded version of work presented at the Late-Breaking Demo session of ISMIR 2019 [Ding et al. 2019]. We thank the Yamaha corporation (particularly R. Tanase, S. Usa, and T. Fujishima) for providing support, feedback, and experiment participants. Additional funding was provided by a Grant-in-Aid from the Japan Society for the Promotion of Science (#19KK0064) and by grants from Keio University (Keio Global Research Institute, Keio Research Institute at SFC, and Keio Gijuku Academic Development Fund) to PES.

9. DATA AVAILABILITY AND SUPPLEMENTARY MATERIAL

Data and analysis code are available at `https://github.com/comp-music-lab/cantometrics`. Audio and Supplementary material is available at `https://osf.io/62vds/`.

10. REFERENCES

Allan, H., Müllensiefen, D., & Wiggins, G. A. (2007). Methodological considerations in studies of musical similarity. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007)*, (pp. 473–478)., Vienna, Austria. Citeseer.

Casey, M. A., Veltkamp, R., Goto, M., Leman, M., Rhodes, C., & Slaney, M. (2008). Content-based music information retrieval: Current directions and future challenges. *Proceedings of the IEEE*, *96*(4), 668–696.

Cheng, D., Joachims, T., & Turnbull, D. (2020). Exploring acoustic similarity for novel music recommendation. *ISMIR 2020*.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213–220.

Daikoku, H., Wood, A. L. C., & Savage, P. E. (2020). Musical diversity in india: A preliminary computational study using cantometrics. *Keio SFC Journal*, *20*(2), 34–61.

Ellis, A. J. (1885). *On the musical scales of various nations*. Journal of the Society of arts.

Flexer, A. & Grill, T. (2016). The problem of limited inter-rater agreement in modelling music similarity. *Journal of new music research*, *45*(3), 239–251.

Grosche, P. & Müller, M. (2011). Tempogram toolbox: Matlab implementations for tempo and pulse analysis of music recordings. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011), Miami, FL, USA*, (pp. 24–28).

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*, 61–135.

Jacoby, N., Margulis, E., Clayton, M., Hannon, E., Honing, H., Iversen, J., Klein, T. R., London, J., Mehr, S., Pearson, L., Peretz, I., Perlman, M., Polak, R., Ravignani, A., Savage, P. E., Steingo, G., Stevens, C., Trainor, L., Trehub, S., Veal, M., & Wald-Fuhrmann, M. (2020). Cross-cultural work in music cognition: Methodologies, pitfalls, and practices. *Music Perception*, *37*(3), 185–195.

Jacoby, N. & McDermott, J. H. (2017). Integer ratio priors on musical rhythm revealed cross-culturally by iterated reproduction. *Current Biology*, *27*(3), 359–370.

Jensen, J. H., Ellis, D. P., Christensen, M. G., & Jensen, S. H. (2007). Evaluation of distance measures between gaussian mixture models of mfccs. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR 2007)*, (pp. 107–108)., Vienna, Austria. Austrian Computer Society.

Jones, M. C., Downie, J. S., & 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 (ISMIR 2007)*, (pp. 539–542)., Vienna, Austria.

Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.

Lee, H., Höger, F., Schönwiesner, M., Park, M., & Jacoby, N. (2021). Cross-cultural mood perception in pop songs and its alignment with mood detection algorithms. In *Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR 2021)*, (pp. 366–373)., Online.

Legendre, P. & Legendre, L. F. (2012). *Numerical ecology*. Oxford: Elsevier.

Light, R. J. (1971). Measures of response agreement for qualitative data: some generalizations and alternatives. *Psychological bulletin*, *76*(5), 365.

List, G. (1974). The reliability of transcription. *Ethnomusicology*, *18*(3), 353–377.

Lomax, A. (1968). *Folk song style and culture*. American Association for the Advancement of Science.

Lomax, A. (1976). *Cantometrics: An approach to the anthropology of music*. Berkeley: Extension Media Center, University of California.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer research*, *27*(2 Part 1), 209–220.

McFee, B., Barrington, L., & Lanckriet, G. (2012). Learning content similarity for music recommendation. *IEEE transactions on audio, speech, and language processing*, *20*(8), 2207–2218.

McFee, B., Barrington, L., & Lanckriet, G. R. G. (2011). Learning content similarity for music recommendation. *IEEE transactions on audio, speech, and language processing*, *20*(8), 2207–2218.

Mehr, S. A., Singh, M., Knox, D., Ketter, D. M., Pickens-Jones, D., Atwood, S., Lucas, C., Jacoby, N., Egner, A. A., Hopkins, E. J., et al. (2019). Universality and diversity in human song. *Science*, *366*(6468), eaax0868.

Nettl, B. & Bohlman, P. V. (1991). *Comparative musicology and anthropology of music: essays on the history of ethnomusicology*. University of Chicago Press.

Panteli, M., Benetos, E., & Dixon, S. (2017). A computational study on outliers in world music. *PlOS ONE*, *12*(12), e0189399.

Panteli, M., Benetos, E., & Dixon, S. (2018). A review of manual and computational approaches for the study of world music corpora. *Journal of New Music Research*, *47*(2), 176–189.

Pauws, S. (2004). Musical key extraction from audio. In *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR 2004)*.

Pohle, T., Schnitzer, D., Schedl, M., Knees, P., & Widmer, G. (2009). On rhythm and general music similarity. In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR 2009)*, (pp. 525–530).

Proutskova, P. (2019). *Investigating the Singing Voice: Quantitative and Qualitative Approaches to Studying Cross-Cultural Vocal Production*. PhD thesis, Goldsmiths, University of London.

Revelle, W. & Condon, D. M. (2019). Reliability from α to ω: A tutorial. *Psychological assessment*, *31*(12), 1395–1411.

Rzeszutek, T., Savage, P. E., & Brown, S. (2012). The structure of cross-cultural musical diversity. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1733), 1606–1612.

Savage, P. E. (2018). Alan lomax's cantometrics project: a comprehensive review. *Music & Science*, *1*, 1–19.

Savage, P. E. & Brown, S. (2013). Toward a new comparative musicology. *Analytical Approaches to World Music*, *2*(2), 148–197.

Savage, P. E., Brown, S., Sakai, E., & Currie, T. E. (2015). Statistical universals reveal the structures and functions of human music. *Proceedings of the National Academy of Sciences of the USA*, *112*(29), 8987–8992.

Savage, P. E., Merritt, E., Rzeszutek, T., & Brown, S. (2012). Cantocore: A new cross-cultural song classification scheme. *Analytical Approaches to World Music*, *2*, 87–137.

Schnitzer, D., Flexer, A., Schedl, M., & Widmer, G. (2011). Using mutual proximity to improve content-based audio similarity. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011)*, volume 11, (pp. 79–84)., Miami, USA.

Sim, J. & Wright, C. C. (2005). The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical therapy*, *85*(3), 257–268.

Tenzer, M. & Roeder, J. (2011). *Analytical and cross-cultural studies in world music*. Oxford University Press, USA.

Tzanetakis, G., Kapur, A., Schloss, W. A., & Wright, M. (2007). Computational ethnomusicology. *Journal of Interdisciplinary Music Studies*, *1*(2), 1–24.

Urbano, J., Bogdanov, D., Herrera, P., Gómez, E., & Serra, X. (2014). What is the effect of audio quality on the robustness of MFCCs and chroma features? In *International Society for Music Information Retrieval Conference (ISMIR 2014)*, (pp. 573–578)., Taipei, Taiwan.

Van Balen, J., Wiering, F., & Veltkamp, R. (2014). Cognitive features for cover song retrieval and analysis. In *Workshop on Folk Music Analysis (FMA2014)*, (pp. 111–113)., Istanbul, Turkey.

Wolff, D. & Weyde, T. (2014). Learning music similarity from relative user ratings. *Information Retrieval*, *17*(2), 109–136.

Wood, A. L. C. (2018). 'Like a cry from the heart': An insider's view of the genesis of Alan Lomax's ideas and the legacy of his research: Part II. *Ethnomusicology*, *62*(3), 403–438.

Wood, A. L. C., Kirby, K. R., Embers, C. R., Silbert, S., Hideo, D., McBride, J., Passmore, S., Paulay, F., Flory, M., Szinger, J., D'Arcangelo, G., Guarino, M., Atayeva, M., Jesse, R., Violet, B., El Hajli, M., Szinger, M., & Savage, P. E. (2021). The global jukebox: A public database of performing arts and culture. *PsyArXiv preprint*.