Agreement among human and automated judgements of similarity in a global music sample

Hideo Daikoku[†1], Shenghao Ding[†1], Ujwal Sriharsha Sanne[2], Emmanouil Benetos[2], Anna Lomax Chairetakis Wood[3], Shinya Fujii[1], Patrick E. Savage[1*]

[1] Keio University, Japan; [2] Centre for Digital Music, Queen Mary University of London, UK; [3]Association for Cultural Equity, USA[†]Equal contribution
[*]Correspondence to: psavage@sfc.keio.ac.jp

© H. Daikoku, S. Ding, et al. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). Attribution:Daikoku, H., Shenghao, D., Sanne, U. S., Benetos, E., Wood, A. L. C., Fujii, S., & Savage, P. E. (2021). Agreement among human and automated judgements of similarity in a global music sample. *PsyArXiv Preprint*.

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

First Author
Affiliation1
author1@ismir.edu

Second Author
Retain these fake authors in submission to preserve the formatting

Third Author
Affiliation3
author3@ismir.edu

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 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 [1]. 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 [2–7]. 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 [8–10]. While various manual systems of cross-cultural classification have been developed, they all have their own drawbacks [2, 3, 10–14]. 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 [15]. Indeed, limited inter-rater agreement has been proposed to represent an upper bound limiting progress in MIR algorithms [16]. Inter-rater agreement has also been raised as an issue potentially limiting cross-cultural musicological analyses [7, 12]. 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 data 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 reviews related work on cross-cultural similarity and inter-rater agreement. Section 3 describes the music dataset used in this study. Section 4 analyzes the inter-rater agreement of our novel perceptual data and compares it with previous studies. Section 5 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.

2. RELATED WORK

2.1 Human judgments of musical similarity

A landmark attempt to measure cross-cultural musical similarity was Alan Lomax's Cantometrics Project (Canto = Song, Metrics = Measure) [11–13, 17]. Lomax and his colleagues analysed thousands of songs from hundreds of

© First Author, Second Author, Third Author. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** First Author, Second Author, Third Author, "Agreement among human and automated judgements of similarity in a global music sample", in *Proc. of the 22nd Int. Society for Music Information Retrieval Conf.,* Online, 2021.

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 the original 5,779 Cantometrics codings and accompanying audio has made it possible to apply it to larger-scale automatic algorithms in the context of music similarity [17]. Recent analysis of the original Cantometrics codings has shown promise in using it to explore musical diversity in India [18]. 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 [7, 12, 19–23]. CantoCore, a classification scheme inspired by Cantometrics, was found to have higher average inter-rater agreement than Cantometrics when compared against a dataset of 30 traditional songs from around the world [19]. Mehr et al. [7] argued that their own scheme was in turn more reliable 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.

Perceptual subjectivity is not only an issue for cross-cultural analyses, but more generally for all studies of musical similarity [15, 16]. Automatic similarity algorithms are generally evaluated against ground-truth human ratings, but when these human ratings were compared against each other in the 2006 Music Information Retrieval Evaluation Exchange (MIREX) evaluations, inter-reliability Kappa scores were found to be relatively limited even for Western popular music [26]. Specifically, when 8 groups of 3 raters each used a 3-point Likert scale of "not similar", "somewhat similar", or "very similar" to rate approximately 200 pairs of songs identified as similar by automated audio similarity algorithms, their average Kappa score was .21-.29 (depending on their approach to weighting partial agreement), which is lower than the average Kappa values of .3-.5 found for Cantometric ratings [17, 19]. Flexer & Grill [16] have argued that limited inter-rater agreement results in upper bounds on the performance of automatic methods of calculating musical similarity.

[1] This comparison is problematic because not only can Cronbach's 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 [7]) gives an inflated appearance of inter-rater agreement compared to smaller numbers of raters (e.g., 2 raters in [19,24]) even if the level of agreement between each pair of raters is the same [25].

2.2 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) [27], tempograms [28], pitch bi-histograms [29] and chromagrams [30] 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 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 [16, 33, 34]. However its accuracy for non-Western music has yet to be systematically evaluated.

Recently, Panteli et al. proposed an automated algorithm [2] 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 [7, 18, 20]. Thus there is still a need to evaluate Panteli et al.'s algorithm against human perceptual data on musical similarity.

3. CROSS-CULTURAL MUSIC DATASET

We chose the 30 audio recordings of traditional songs from Lomax's Cantometrics Consensus Tape [11] 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 to classify songs around the world, and thus already functioned as one type of pre-annotated expert ground-truth data [19, 35]. Not only were

[2] MIREX also evaluates symbolic music similarity as represented through staff notation. However, the reliability of such notation for non-Western music is debated [31, 32] and has not yet been objectively evaluated, so we have not included symbolic notation in the current study.

[3] Flexer and Grill [16] explain that this section was discontinued from MIREX from 2015 because "only our own research team, again sending the same peak performing system PS2 [33] for the seventh year, wanted to participate".

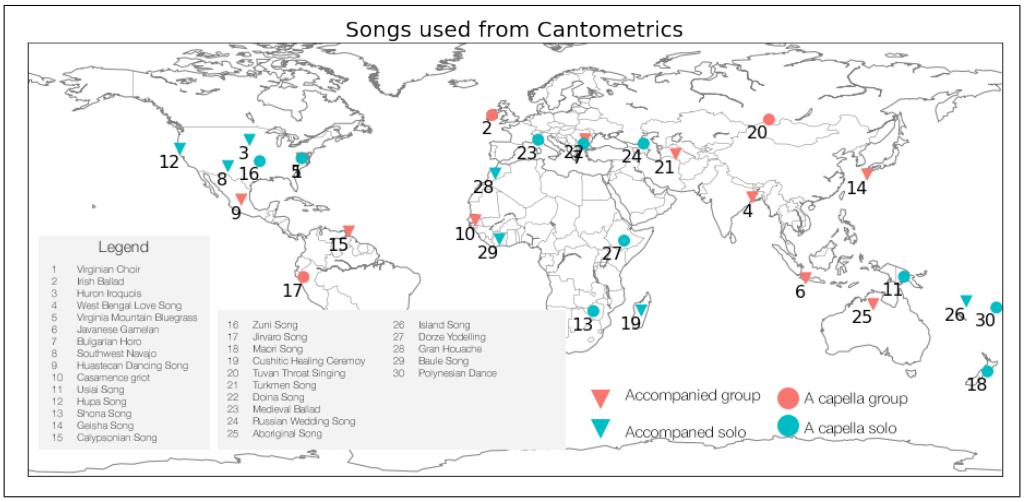

Figure 1. A map showing the approximate locations and cultural group names of the 30-song sample from the Cantometrics Consensus Tape (adapted from [19, 35]). We label songs by solo vs. group singing and with vs. without instrumental accompaniment.

these 30 songs pre-annotated using Cantometrics, but almost 6,000 more songs have also been annotated using Cantometrics [11–13, 17], 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 4).

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 [19]. 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 [18, 24]. 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.

4. PERCEPTUAL EXPERIMENT

4.1 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

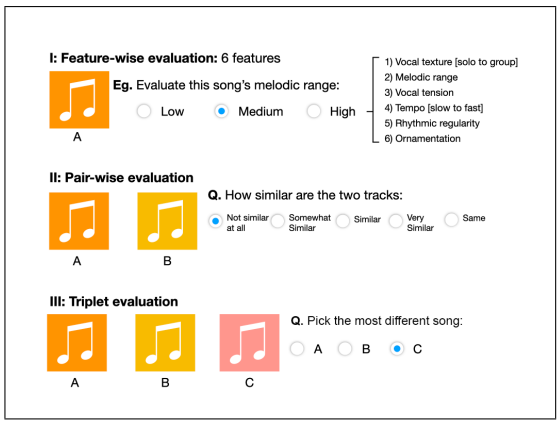

Figure 2. The experiment is divided into three parts. Part I asks participants to rate each song on a 3-point scale for 6 stylistic features. Part II asks the rater to listen to two songs and rate how similar they sound on a 5-point scale. Part III asks raters to listen to three samples and pick the most different song.

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

4.1.1 Songwise Evaluation

Part I of our experiment required participants to rate 5 songs individually on the basis of six stylistic features[4] : vocal texture, melodic range, vocal tension, tempo, rhythmic regularity, and ornamentation. Some features were relabeled to make them more intuitive in the context of the current experiment; the original Cantometric feature names and line numbers were "The vocal group" (Line 1), "Melodic range" (20), "Vocal width" (33), "Tempo" (26), "Overall vocal rhythmic scheme" (11), and "Embellishment" (23); see Figure 2. These features were selected as a subset of the 37 Cantometrics features, chosen in order to focus on features that were likely to be relatively reliable while covering a diverse range of musical features (e.g., melody, rhythm, timbre) and minimizing correlations among features (see [12] for an overview of the full 37 features and their inter-relationships). Before starting the experiment, participants spent approximately 5 minutes listening to reference examples from the Cantometrics Training Tapes demonstrating low, medium and high values for each feature, and participants could refer back to these reference examples at any point during the experiment.

[4] Ratings for an additional 7 aesthetic features were also collected for future analyses comparing stylistic and aesthetic features; these features are shown in Figure S1 for completeness but are not included in our analyses.

4.1.2 Pairwise Comparison

In part II of our experiment we ask raters to listen to pairs of songs and rate their similarity on a five-point Likert scale. Participants were asked to rate each pair as "not at all similar", "a little similar", "somewhat similar", "very similar", and "extremely similar". We compare all 10 possible pairs among each set of 5 songs provided to the group, with the order randomized for each participant.

4.1.3 Triplet Comparison

In part III we ask participants to listen to three songs and pick the most dissimilar song. We test for all 10 possible triplet combinations among the set of five songs and the order is randomized.

4.2 Participants

We tested 62 participants employed at X[5] '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. 73% of participants had little to no exposure to world music, and 92% of participants listened to a combination of rock, pop, jazz or classical music.

4.3 Human-human agreement

We calculated Kappa statistics to determine inter-rater agreement of our perceptual data. 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 [36, 37]. 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 cf. Figure 3).

Based on previous research [38] we predicted that triplet ratings would be more reliable than pairwise or feature-wise ratings. Surprisingly, we found the opposite was true (mean Kappas: pairwise = 0.29, triplet = 0.19, featurewise = 0.30; F = 11.78, df= 2377, p=0.000008). These patterns were consistent when data were re-analyzed with participants split into subsets with high (>2 years) or low (<2 years) musical training (Supplementary Figure S1). Supplementary Figure S2 compares feature wise inter-rater reliability from this study with feature wise reliability of previous studies [7, 19, 24].

The mean level of inter-rater agreement among untrained participants in the current study was .30, a level that

[5] Anonymized for peer review

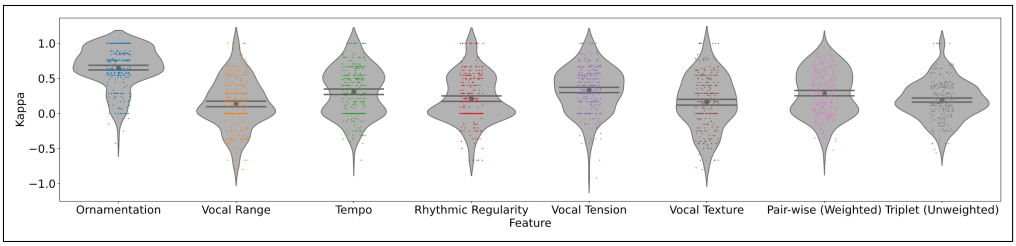

Figure 3. Violin plots showing inter-rater reliability (Kappa) for the different rating methods used in the current study: pairwise, triplet, and feature wise for six different stylistic features. Each individual point ($n = 290 - 300$ per condition) represents a Kappa value for a given pair of participants.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

has been described as "fair" [39]. This is toward the low end of the range of 0.25-0.54 found previously for experts using Cantometrics and similar schemes [7, 17, 19, 24].

5. 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 [34]; A2: Panteli et al. [2] ("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 (because it was not practical to collect perceptual ratings of pairwise similarity for all 435 pairs among the 30 songs for each participant).

5.1 Distance matrix correlations

Because H2 could not include the full 30x30 distance matrix, we performed two sets of distance matrix correlations (Figure 4). 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 [40]. 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 [41].

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

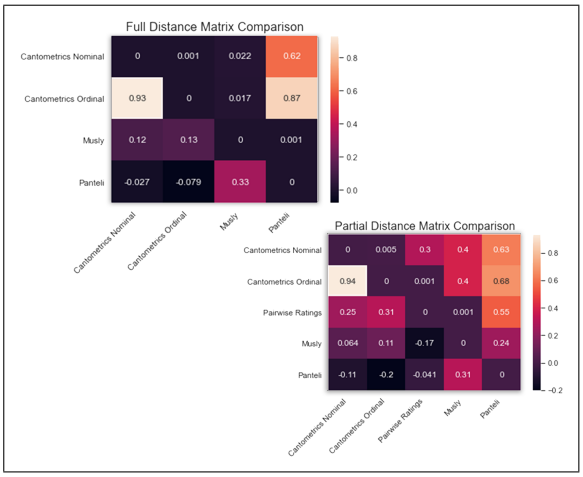

Figure 4. Distance matrix correlations among different methods of measuring musical similarity for the full 30x30 distance matrix and the partial distance matrix containing six 5x5 distance matrices. The bottom left triangle indicates the correlation coefficient, and the top right triangle shows its corresponding p-value.

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).

6. DISCUSSION

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 [5]. 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 [16]. 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) [26]. These levels are significantly above chance, but not particularly high (some have suggested .21-0.4 is 'fair' [39], while others argue less than 0.4 is unacceptable for certain applications (e.g., clinical diagnoses) [42]. 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.

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 [38, 43] 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 [20]).

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 [19]. This suggests that there is nothing inherently insurmountable about cross-cultural comparison of music [10], 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.

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 [11–13, 17]. 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 [43–45]. 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.

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 [46] towards homogeneous "WEIRD" (Western, Educated, Industrialized, Rich, Democratic [47]) 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 [46, 48].

The two audio similarity algorithms evaluated in this study were chosen because they either performed best in MIREX's audio similarity evaluation, or in [2]'s case were designed specifically for measuring similarity in non-Western music. However, these systems have limitations, such as emphasis on timbral features [34] and reliance on country labels as a proxy for similarity [2] that may contribute to their poor performance. In the future we would like to explore alternative methods such as existing Gaussian mixture models [45], metric learning [43, 49], and deep learning [44]. 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 [17].

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

[Omitted for double-blind review]

8. ACKNOWLEDGMENTS

[Omitted for double-blind review] This study was approved by the [omitted] University Research Ethics Committee (approval No. 211).

9. DATA AVAILABILITY

Data and analysis code are available at `https://drive.google.com/file/d/1I01rcTGki1_l7gbtY3C-8tN2rj6TMyNc/view?usp=sharing` [temporary link to anonymized GitHub repository clone for double-blind review]. The full experiment can be accessed at `https://[anonymizedfordouble-blindreview]`.

10. REFERENCES

[1] M. A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, and M. Slaney, "Content-based music information retrieval: Current directions and future challenges," *Proceedings of the IEEE*, vol. 96, no. 4, pp. 668–696, 2008.

[2] M. Panteli, E. Benetos, and S. Dixon, "A computational study on outliers in world music," *PlOS ONE*, vol. 12, no. 12, p. e0189399, 2017.

[3] ——, "A review of manual and computational approaches for the study of world music corpora," *Journal of New Music Research*, vol. 47, no. 2, pp. 176–189, 2018.

[4] P. E. Savage, "An overview of cross-cultural music corpus studies," in *Oxford Handbook of Music and Corpus Studies*, D. Shanahan, A. Burgoyne, and I. Quinn, Eds. New York: Oxford University Press, In press.

[5] G. Tzanetakis, A. Kapur, W. A. Schloss, and M. Wright, "Computational ethnomusicology," *Journal of Interdisciplinary Music Studies*, vol. 1, no. 2, pp. 1–24, 2007.

[6] J. Six, O. Cornelis, and M. Leman, "Tarsos, a modular platform for precise pitch analysis of western and non-western music," *Journal of New Music Research*, vol. 42, no. 2, pp. 113–129, 2013.

[7] S. A. Mehr, M. Singh, D. Knox, D. M. Ketter, D. Pickens-Jones, S. Atwood, C. Lucas, N. Jacoby, A. A. Egner, E. J. Hopkins *et al.*, "Universality and diversity in human song," *Science*, vol. 366, no. 6468, p. eaax0868, 2019.

[8] A. J. Ellis, *On the musical scales of various nations*. Journal of the Society of arts, 1885.

[9] B. N. P. V. Bohlman, *Comparative musicology and anthropology of music: essays on the history of ethnomusicology*. University of Chicago Press, 1991.

[10] P. E. Savage and S. Brown, "Toward a new comparative musicology," *Analytical Approaches to World Music*, vol. 2, no. 2, pp. 148–197, 2013.

[11] A. Lomax *et al.*, *Folk song style and culture*. American Association for the Advancement of Science, 1968.

[12] P. E. Savage, "Alan lomax's cantometrics project: a comprehensive review," *Music & Science*, vol. 1, pp. 1–19, 2018.

[13] A. L. C. Wood, ""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*, vol. 62, no. 3, pp. 403–438, 2018. [Online]. Available: https://www.jstor.org/stable/10.5406/ethnomusicology.62.3.0403

[14] M. Tenzer and J. Roeder, *Analytical and cross-cultural studies in world music*. Oxford University Press, 2011.

[15] H. Allan, D. Müllensiefen, and G. A. Wiggins, "Methodological considerations in studies of musical similarity." in *ISMIR*. Citeseer, 2007, pp. 473–478.

[16] A. Flexer and T. Grill, "The problem of limited inter-rater agreement in modelling music similarity," *Journal of new music research*, vol. 45, no. 3, pp. 239–251, 2016.

[17] A. L. C. Wood, K. R. Kirby, C. R. Embers, S. Silbert, D. Hideo, J. McBride, S. Passmore, F. Paulay, M. Flory, J. Szinger, G. D'Arcangelo, M. Guarino, M. Atayeva, R. Jesse, B. Violet, M. El Hajli, M. Szinger, and P. E. Savage, "The global jukebox: A public database of performing arts and culture," *PsyArXiv preprint*, 2021. [Online]. Available: https://doi.org/10.31234/osf.io/4z97j

[18] H. Daikoku, A. L. C. Wood, and P. E. Savage, "Musical diversity in india: A preliminary computational study using cantometrics." *Keio SFC Journal*, vol. 20, no. 2, pp. 34–61, 2021. [Online]. Available: https://gakkai.sfc.keio.ac.jp/journal/en/210325e.html

[19] P. E. Savage, E. Merritt, T. Rzeszutek, and S. Brown, "Cantocore: A new cross-cultural song classification scheme." *Analytical Approaches to World Music*, 2012.

[20] T. Rzeszutek, P. E. Savage, and S. Brown, "The structure of cross-cultural musical diversity," *Proceedings of the Royal Society B: Biological Sciences*, vol. 279, no. 1733, pp. 1606–1612, 2012.

[21] P. Proutskova and M. Casey, "You call THAT singing? Ensemble classification for multi-cultural collections of music recordings," in *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR 2009)*, 2009, pp. 759–764.

[22] P. Proutskova, "Investigating the singing voice: Quantitative and qualitative approaches to studying cross-cultural vocal production," Ph.D. dissertation, Goldsmiths, University of London, February 2019. [Online]. Available: http://research.gold.ac.uk/26133/

[23] M. Magas and P. Proutskova, "A location-tracking interface for ethnomusicological collections," *Journal of New Music Research*, vol. 42, no. 2, pp. 151–160, Jun. 2013. [Online]. Available: https://doi.org/10.1080/09298215.2013.821999

[24] P. E. Savage, S. Brown, E. Sakai, and T. E. Currie, "Statistical universals reveal the structures and functions of human music," *Proceedings of the National Academy of Sciences of the USA*, vol. 112, no. 29, pp. 8987–8992, 2015.

[25] W. Revelle and D. M. Condon, "Reliability from α to ω: A tutorial." *Psychological assessment*, vol. 31, no. 12, pp. 1395–1411, 2019.

[26] M. C. Jones, J. S. Downie, and A. F. Ehmann, "Human similarity judgments: Implications for the design of formal evaluations." in *ISMIR*, 2007, pp. 539–542.

[27] J. Urbano, D. Bogdanov, P. Herrera, E. Gómez, and X. Serra, "What is the effect of audio quality on the robustness of MFCCs and chroma features?" in *International Society for Music Information Retrieval Conference*, 2014, pp. 573–578.

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

[29] J. Van Balen, F. Wiering, and R. Veltkamp, "Cognitive features for cover song retrieval and analysis," in *Workshop on Folk Music Analysis (FMA2014)*, 2014, pp. 111–113.

[30] S. Pauws, "Musical key extraction from audio." in *ISMIR*, 2004.

[31] G. List, "The reliability of transcription," *Ethnomusicology*, vol. 18, no. 3, pp. 353–377, 1974.

[32] N. M. England, R. Garfias, M. Kolinski, G. List, W. Rhodes, and C. Seeger, "Symposium on transcription and analysis: A Hukwe* song with musical bow," *Ethnomusicology*, vol. 8, no. 3, pp. 223–233, 1964.

[33] T. Pohle, D. Schnitzer, M. Schedl, P. Knees, and G. Widmer, "On rhythm and general music similarity." in *ISMIR*, 2009, pp. 525–530.

[34] D. Schnitzer, A. Flexer, M. Schedl, and G. Widmer, "Using mutual proximity to improve content-based audio similarity." in *ISMIR*, vol. 11, 2011, pp. 79–84.

[35] A. Lomax, *Cantometrics: An approach to the anthropology of music*. Extension Media Center, University of California, 1976.

[36] J. Cohen, "Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit," *Psychological Bulletin*, vol. 70, no. 4, pp. 213–220, 1968.

[37] R. J. Light, "Measures of response agreement for qualitative data: some generalizations and alternatives." *Psychological bulletin*, vol. 76, no. 5, p. 365, 1971.

[38] D. Wolff and T. Weyde, "Learning music similarity from relative user ratings," *Information Retrieval*, vol. 17, no. 2, pp. 109–136, 2014.

[39] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.

[40] P. Legendre and L. F. Legendre, *Numerical ecology*. Elsevier, 2012.

[41] N. Mantel, "The detection of disease clustering and a generalized regression approach," *Cancer research*, vol. 27, no. 2 Part 1, pp. 209–220, 1967.

[42] J. Sim and C. C. Wright, "The kappa statistic in reliability studies: use, interpretation, and sample size requirements," *Physical therapy*, vol. 85, no. 3, pp. 257–268, 2005.

[43] B. McFee, L. Barrington, and G. Lanckriet, "Learning content similarity for music recommendation," *IEEE transactions on audio, speech, and language processing*, vol. 20, no. 8, pp. 2207–2218, 2012.

[44] D. Cheng, T. Joachims, and D. Turnbull, "Exploring acoustic similarity for novel music recommendation," *Currently Under Review*, 2020.

[45] J. H. Jensen, D. P. Ellis, M. G. Christensen, and S. H. Jensen, "Evaluation of distance measures between gaussian mixture models of mfccs," in *Proceedings of the 8th International Conference on Music Information Retrieval*. Austrian Computer Society, 2007, pp. 107–108.

[46] N. Jacoby, E. Margulis, M. Clayton, E. Hannon, H. Honing, J. Iversen, T. R. Klein, J. London, S. Mehr, L. Pearson, I. Peretz, M. Perlman, R. Polak, A. Ravignani, P. E. Savage, G. Steingo, C. Stevens, L. Trainor, S. Trehub, M. Veal, and M. Wald-Fuhrmann, "Cross-cultural work in music cognition: Methodologies, pitfalls, and practices," *Music Perception*, vol. 37, no. 3, pp. 185–195, 2020.

[47] J. Henrich, S. J. Heine, and A. Norenzayan, "The weirdest people in the world?" *Behavioral and Brain Sciences*, vol. 33, pp. 61–135, 2010.

[48] N. Jacoby and J. H. McDermott, "Integer ratio priors on musical rhythm revealed cross-culturally by iterated reproduction," *Current Biology*, vol. 27, no. 3, pp. 359–370, 2017.

[49] B. McFee, L. Barrington, and G. R. G. Lanckriet, "Learning content similarity for music recommendation," *CoRR*, vol. abs/1105.2344, 2011. [Online]. Available: http://arxiv.org/abs/1105.2344

[50] J. Salamon and E. Gómez, "Melody extraction from polyphonic music signals using pitch contour characteristics," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, pp. 1759–1770, 08/2012 2012. [Online]. Available: http://hdl.handle.net/10230/42183

[51] E. M. Von Hornbostel and C. Sachs, "Systematik der musikinstrumente. ein versuch," *Zeitschrift für Ethnologie*, vol. 46, no. H. 4/5, pp. 553–590, 1914.

11. SUPPLEMENTARY MATERIALS

11.1 Creating distance matrices

11.1.1 H1: Expert (Cantometrics)

Our 30-song sample was chosen in part because it already contained expert annotations supplied by Alan Lomax using the Cantometrics classification scheme [35]. Each of the 36 Cantometric features was rated on a Likert-type scale, the number of points of which ranged from 3 to 13 [11, 12]. Some of these features were clearly ordinal (ordered) - for example, the presence of tremolo (vocal micro–ornamentation including vibrato) is coded on a 3-point ordinal scale of "little or no tremolo", "some tremolo", or "much tremolo". Other features were intended to be ordinal, but could be argued to be more appropriately treated as nominal (unordered) because the order was semi-arbitrary. For example, melodic shape was coded on a 4-point scale of "arched", "terraced", "undulating", and "descending", but it is not obvious whether an undulating shape is more similar to a descending shape than to an arched shape. Another example of semi-arbitrary order is vocal organization, which is also coded on a 4-point scale of "solo", "unison", "heterophony", and "polyphony". To further complicate matters, Cantometrics allows features to be multi-coded - for example songs with heterogeneous sections can be rated as containing a combination of "arched" and "descending" phrase shapes or as having both "little or no tremolo" and "heavy tremolo".

To convert these annotations into quantitative distance matrices, we adapted the method previously used to calculate similarities using Cantometrics and the related CantoCore scheme [10, 20, 24]. This method normalizes ordinal variables and accounts for partial similarities between multi-coded ordinal variables to create a multi-dimensional Euclidean distance matrix ranging from 0 (songs with identically coded features) to 1 (songs with maximally contrasting features) [20]. We had to adapt this method slightly because, unlike CantoCore, Cantometrics allows multi-coding for both nominal and ordinal variables. Thus, to accommodate double-coding for ordinal variables, mean values are used in cases of multi-coded ordinal features (e.g., a song in which tremolo was coded as both 2 [some] and 1 [much] would be treated as 1.5).

To account for ambiguity in whether some features should be treated as ordinal vs. nominal, we calculated two separate Cantometric distance matrices. The first, which we call "Cantometrics nominal" (H1a) treated 12 of the 36 features (numbers 1-4, 7, 11-16, and 22) as nominal, and the remaining 24 features as ordinal. The second ("Cantometrics ordinal"; H1b) treated all 36 variables as ordinal.

11.1.2 H2a: Naive listener (5x5 pairwise ratings x 6)

Distance matrices of naive listener ratings were taken directly from the pairwise similarity ratings in Section 4.1.2. Similarity ratings for each pair and each participant were scaled from 0 to 1 and converted from similarities to distances by subtracting from 1. Therefore, "not at all similar" was treated as a distance of 1, "a little similar" = 0.75, "somewhat similar" = 0.5, "very similar" = 0.25, and "extremely similar" = 0. Ratings for all 10 pairs within each set of 5 songs were then averaged for across all 10-11 participants who rated those songs.

11.1.3 A1: Musly

Musly is an open source music similarity algorithm implemented using timbre as the principal axis for proximity [34]. We used the default implementation, which uses a 25 MFCC representation of each song to estimate a single Gaussian where the similarities are normalized with Mutual Proximity. We add the playlist of 30 songs as input and then compute the full similarity matrix for the specified collection and write it to a file written in the MIREX text format. All implementation details are the same as publicly available online [34].

11.1.4 A2: Panteli

We selected the LDA (Linear Discriminant Analysis) implementation of Panteli et al.'s algorithm because it provided the most reliable classifications, as measured by F-score [2]. The algorithm first extracts MFFCs, chromagrams, pitch bi-histograms, and tempograms from a dataset of 8200 recordings from the Smithsonian Folkways Recordings and British Library Sound Archive. The model also uses the Melodia Vamp plugin [50] for Sonic Annotator for melody extraction and a Song/Speech segmenter for removing parts of speech from the original sample (although this was not necessary for the current sample because it contained only music). All sampling is at 44100 Hz using 8-second windows with 0.5-second hop size. It then pools the data into an 840 dimension vector, along with country labels as training data. This is set as input data for an LDA model that computes dissimilarity of a given input vector of the same dimensions. The reason for choosing optimal parameter values and other implementation details are described in the original paper [2]. We perform the same feature extraction as that implemented in the algorithm on our input data of 30 songs and compute a distance matrix.

11.2 Comparing inter-rater agreement with previous studies

Figure S2 compares inter-rater agreement in the current study with previous studies. Inter-rater agreement for 36 Cantometric features and 26 CantoCore features was previously reported for two raters [19]. Reliability for an adapted scheme taking 32 putative universal features from CantoCore, Cantometrics, and the Hornbostel-Sachs instrument classification schemes [51] and condensing them into binary presence/absence features was also previously reported [24]. Mehr et al. [7] did not report Kappa data, but using their publicly available data and code we were able to calculate Kappa statistics for 15 of their 16 features reported in their Table S31 using the same methods reported here (Mehr et al.'s "tempo_adj" feature was excluded because it was neither ordinal nor nominal but continuous).

We applied quadratic weighted kappa on Mehr et al.'s ordinal features (macro meter, syncopation, accent, tension, and scale quality), and unweighted kappa on other binary features.

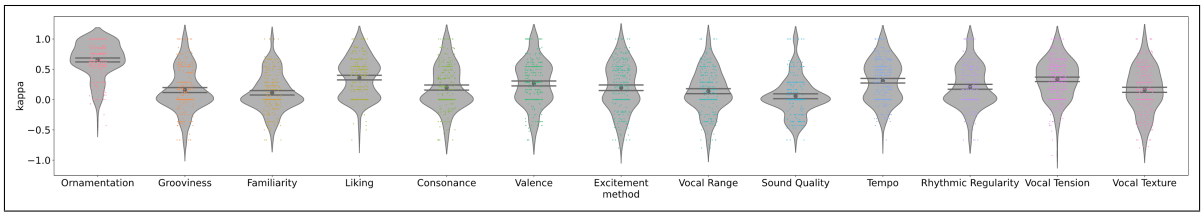

Figure S1. An extended version of Figure 3 that also includes inter-rater reliability data for 7 additional aesthetic features that were collected but not analyzed in the current study.

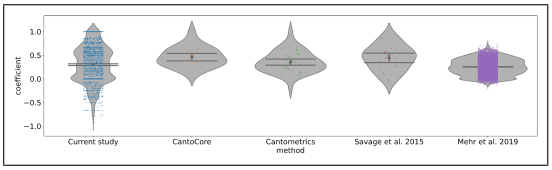

Figure S2. Comparison of inter-rater reliability for individual features in the current study compared to previous studies. Each point represents a Kappa value for a given pair of raters for a given musical feature. Sample sizes: Cantometrics: $n = 36$; CantoCore: $n = 26$; Savage et al. 2015: $n = 32$; Mehr et al. 2019: $n = 6525$ (435 pairs of raters x 15 features); current study: $n = 1788$ (298 pairs of raters x 6 features). Large dots represent means, horizontal lines represent 95% confidence intervals.

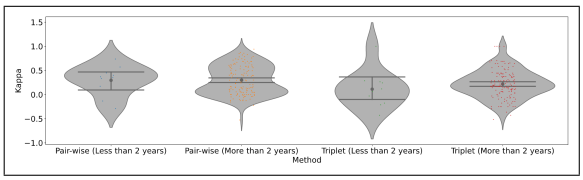

Figure S3. Inter-rater reliability analyses for subsets of raters with high (>2 years) and low (<2) levels fo musical training.