AGREEMENT AMONG HUMAN AND AUTOMATED TRANSCRIPTIONS OF GLOBAL SONGS

Yuto Ozaki1, John McBride2, Emmanouil Benetos3, Peter Q. Pfordresher4, Joren Six5, Adam T. Tierney6, Polina Proutskova3, Emi Sakai7, Haruka Kondo1, Haruno Fukatsu1, Shinya Fujii1, Patrick E. Savage4*

1Keio University, Fujisawa, Japan
2Center for Soft and Living Matter, Institute for Basic Science, South Korea
3Queen Mary University of London, UK
4University at Buffalo, NY, USA
5Ghent University, Belgium
6Birkbeck, University of London, UK
7No affiliation
8psavage@sfc.keio.ac.jp

ABSTRACT
Cross-cultural musical analysis requires standardized symbolic representation of sounds such as score notation. However, transcription into notation is usually conducted manually by ear, which is time-consuming and subjective. Our aim is to evaluate the reliability of existing methods for transcribing songs from diverse societies. We had 3 experts independently transcribe a sample of 32 excerpts of traditional monophonic songs from around the world (half a cappella, half with instrumental accompaniment). 16 songs also had pre-existing transcriptions created by 3 different experts. We compared these human transcriptions against one another and against 10 automatic music transcription algorithms. We found that human transcriptions can be sufficiently reliable (~90% agreement, κ ~.7), but current automated methods are not (~60% agreement, κ <.4). No automated method clearly outperformed others, in contrast to our predictions. These results suggest that improving automated methods for cross-cultural music transcription is critical for diversifying MIR.

1. INTRODUCTION
Cross-cultural analysis is essential to explore diversity and universality of music [1-2]. Such analyses require symbolic representations of sounds such as score notation. However, transcription into notation is usually conducted by ear, which is time-consuming and subjective [3-4]. Automated methods of music transcription and melody extraction might potentially solve these problems [5-7]. However, automated extraction of fundamental frequency (F0) alone is not sufficient. Instead, a continuous fundamental frequency must be segmented into discrete notes with the categorical pitches and rhythms that are distinctive features of almost all the world’s music [8]. This challenge is particularly important for variable pitch instruments such as the voice (the most universal instrument [8-9]). However, to our knowledge, agreement among human and automated transcription has not been objectively quantified using cross-cultural samples or multiple human transcribers.

The main objective of this paper is to evaluate the degree of agreement among human and automated transcriptions for a global song sample. We demonstrate that the degree of agreement between human transcriptions is substantially higher than the agreement between humans and machines. Our evaluation also reveals that no single algorithm outperforms the others, and there are no clear differences between signal-processing-based methods and data-driven methods.

2. RELATED WORK
2.1 Subjectivity of manual transcription
Manual transcription is central to musicological research, but to our knowledge, agreement among different human transcriptions of the same songs has never been objectively measured. Even qualitative evaluation is rare. A notable exception was a 1963 symposium on transcription where four leading ethnomusicologists independently transcribed a single recording (“A Hukwe* song with musical bow”), resulting in “four rather different transcriptions” [1, 4]. In contrast, List compared transcriptions of three songs (“Romanian carol”, “Yiddish lullaby”, “Thai lullaby”) by between 2-9 transcribers and concluded that “transcriptions made by ear in notated form are sufficiently accurate, sufficiently reliable to provide a valid basis for analysis” [3]. More recently, Mehr et al. [9] combined transcriptions by 3 experts of 118 diverse traditional songs into a single set of “consensus” transcriptions, and had 10 experts rate their accuracy on a subjective scale from 1 (“Terrible”) to 8 (“Perfect”), finding a median rating of 6 (“Very accurate”). Yet none of these studies provided an objective measurement of the degree of agreement between individual transcribers.
2.2 Reliability of automated transcription

Automatic transcription reliability has been evaluated extensively for piano music and some other genres of Western music, but rarely for non-Western music. Ycart et al. [10] evaluated the performance of four automated transcription systems against perceptual ratings from 186 participants over 153 examples of piano music taken from the MAPS dataset of MIDI-aligned piano recordings [11]. They found an average Fleiss’ Kappa coefficient of 0.59, or “borderline between moderate and substantial agreement” on participant ratings. Holzapfel and Benetos asked 16 musicologists from 3 European universities to transcribe 8 excerpts of sousta, a traditional Greek instrumental dance genre, either from scratch or starting from an automatic transcription, finding “no quantitative advantage of using [automatic music transcription]” [12]. Although computer-assisted transcription studies exist [13], recent reviews by musicologists argued that computational tools for musical analysis are either useful for only low-level analysis or not widely adopted within mainstream musicology [14-15]. Overall, there is a clear need for objective measurements of agreement among automated and human transcriptions for a cross-cultural sample of songs.

3. METHODS AND DATASET

3.1 Audio data

To examine the degree of agreement among human and automated transcriptions of diverse songs, we collected a sample of 14-second excerpts from 32 traditional songs evenly distributed across 8 geographic regions (Fig. 1). 16 songs were sampled from the publicly available 14-second excerpts of the Natural History of Song (NHS) Discography dataset [9] and manually extracted 14-second excerpts of the Global Jukebox audio files [16], respectively. We choose these datasets since they cover traditional songs from a global sample of societies. Sampling is randomly conducted using the following criteria:

- Songs are sampled equally from each of the eight regions previously used by NHS for their sampling (i.e. 4 songs per region from North America, Oceania, etc.).
- To assess capabilities of extracting vocal melodies from instrumental accompaniment, songs are sampled to consist of half solo singing without instruments and half solo singing with instruments.

One exception is that the NHS dataset contains no audio recordings of solo singing with instruments in the Middle East region, so two solo singing excerpts without instrument examples were chosen from this region instead. We deliberately let sampled audio recordings contain various degrees of noise, reflecting the real-world challenges of analyzing traditional recordings. We did not include songs with polyphonic singing since polyphonic transcription is substantially more challenging for both humans and automated methods [5], and is beyond the scope of this study.

3.2 Automated methods

We selected 10 automatic music transcription/vocal melody extraction/pitch detection methods. We first choose methods listed in [10] as a baseline. However, that study focused on the systems designed for piano music, so we add methods designed for extracting pitch from human vocals. Considering the difference in the approach of the pitch estimation, our selection consists of automated methods from non-data-driven models and data-driven models. If the model employs a machine learning method (such as artificial neural networks) to learn model parameters from data in a training step, we call it data-driven, otherwise non-data-driven. Table 1 summarizes the selected automated methods. Regarding pYIN [17], we used the TONY [18] software to obtain its F0 estimation. Recently, several symbolic-level automatic transcription methods have been developed [19-21]. However, some models were evaluated with only MIDI synthesized sounds and were not specifically designed for singing voice, so we did not select those methods.

3.3 Transcription process

Figure 1. Map of the 32 songs transcribed and analyzed.
Twelve-tone equal temperament (12-TET) with A4 = 440 Hz is used to transcribe audio into staff notation by humans. Equal temperament is also applied to automated methods to standardize their outputs. As explained in the introduction, it is essential to obtain symbolic representations of pitch contours to analyze acoustic stimuli as melody. However, 12-TET is not completely appropriate since the pitch quantized into 12-TET does not always correspond to the actual scales/modes and perceptual tonal models even for Western singing, let alone non-Western [30-31].

<table>
<thead>
<tr>
<th>Method</th>
<th>Target sound</th>
<th>Unit</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Melodia [22]</td>
<td>Vocal melody</td>
<td>Frame</td>
<td>Data-driven: parameters optimized by training with datasets.</td>
</tr>
<tr>
<td>SS-nPNN [26]</td>
<td>Vocal melody</td>
<td>Frame</td>
<td>Data-driven: parameters optimized by training with datasets.</td>
</tr>
<tr>
<td>AD-NPPN [27]</td>
<td>Multiple piano sound</td>
<td>Note</td>
<td>Data-driven: parameters optimized by training with datasets.</td>
</tr>
<tr>
<td>OAF [28]</td>
<td>Multiple piano sound</td>
<td>Note</td>
<td>Data-driven: parameters optimized by training with datasets.</td>
</tr>
<tr>
<td>madmom [29]</td>
<td>Multiple piano sound</td>
<td>Note</td>
<td>Data-driven: parameters optimized by training with datasets.</td>
</tr>
</tbody>
</table>

Table 1. Summary of the selected automated methods. Unit indicates if the F0 estimation is frame-level or note-level [5] that the latter predicts onset and offset timing.

While binning continuous F0 into a simplified discrete set of 12 100-cent intervals loses information about microtonal nuance, 100 cents (1 semitone) is both the most commonly used system and roughly corresponds to general levels of variability in singing intonation (imprecision and inaccuracy) [31-32], making it a reasonable choice to use to evaluate accuracy. It's also what was used by Mehr et al. [9] when creating the dataset we use, enabling us to compare our results with theirs. In summary, we decide to take advantage of the convenience and comparability of 12-TET, while acknowledging that it does not capture all musical nuances.

This study focuses on the evaluation of agreement among melodies, and we discard temporal/rhythmic information so we only extract pitch from transcriptions to create a sequence of notes. However, regarding the notes representing unison melodic intervals (i.e., repeated notes), we create two transcription patterns. This is because not all selected automated methods can perform note segmentation. The change in pitch class can be used to segment two notes in the case of the other intervals, but the determining boundary between the notes of the same pitch class would require a note segmentation algorithm.

Firstly, the raw note sequences are created as a note sequence which includes the unison interval. Based on this version, we also create a note sequence which discards repeated notes and treats the notes of the unison interval as a tied single note (i.e., “CCFGGC” becomes “CFGC”). We call this version “non-unison”. This treatment enables us to evaluate how much the pitch estimation itself, which is a baseline function of automatic transcription, determines performance. In addition, 12-TET has enharmonic equivalent pitch classes, so we only use flat notes for the same sounding sharp and flat notes.

3.3.1 Transcription by humans

We asked three Japanese experts with professional training in Western classical music to independently transcribe the 32 recordings. One of them has professional experience of transcribing non-Western music using Western staff notation. None of them had seen the transcriptions contained in the NHS dataset. They were instructed to use MuseScore3 [33] as a tool to create transcriptions. Following Mehr et al. [9]1, we also created a consensus version of our 3 new human transcriptions. Importantly, however, while Mehr et al. only analyzed and published their consensus transcription, we include the three independent transcriptions as well as their combined consensus version to allow us to measure agreement between individual human transcribers. Our three coauthors who undertook transcription were blinded from our hypothesis testing and were asked to create transcriptions prior to discussions about coauthorship.

3.3.2 Transcription by automated methods

In order to standardize the output of each method, we apply post-processing steps including manual work, such as the quantization of frequency, smoothing of pitch contour, or the selection of melody contour by the Viterbi algorithm with manually specifying frequency range of melody for the case of multi-pitch estimation methods (cf. supplementary materials for details). Note that songs used in the evaluation contain solo singing with instrumental accompaniment but chosen methods are not designed to estimate the F0 of those styles of singing except for Melodia and SS-nPNN. Therefore the automated methods other than Melodia and SS-nPNN may include the pitch estimation of instrumental sounds, which is excluded from human transcriptions.

3.4 Sequence alignment and evaluation metrics

We use the Needleman-Wunsch algorithm [34] to align note sequences (cf. supplementary materials for further details). Agreement between two string sequences can be quantified in various ways. We mainly use Fleis' Kappa inter-rater reliability coefficient (κ), which measures how much the observed agreement exceeds chance [35]. However, κ does not provide other relevant information such as how many notes actually differ among note sequences or whether differences are due to disagreement about note

1 Mehr et al.'s full consensus transcriptions are published at https://osf.io/jh7is/.
Figure 2. Overview of the agreement evaluation using an example 8-second excerpt from NAIV-075 (Healing song, Kwakwaka’wakw people, 00:06-00:14 from https://osf.io/y29wp). Red indicates disagreement with our new consensus transcription (#4, made by combining the three individual transcriptions #1-3). For visibility, only the automated transcription produced by TONY is shown, and octave information is omitted from the note sequences. The degree of human-human and human-machine agreement is calculated based on the note sequences (c). For example, %5 (NHS consensus) is 95% identical to our consensus #4 (14 out of 15 notes each), while TONY is only 48% identical (7 identical notes out of average note length of 14.5), corresponding to Fleiss’ Kappa values of .94 and .34, respectively. *NB: NHS consensus transcriptions were not available for the 16 songs from the Global Jukebox sample.

3.5 Transposition

We applied transposition in the note sequence alignment process to exclude the effect of disagreement by the discrepancy of the key when calculating $\kappa$, PID and Levenshtein distance. The transposition interval was searched from -2 semitones to +2 semitones. For human-human transcription comparison, the transposition was performed to maximize PID. Regarding the human-machine transcription comparison, the transposition interval was searched to maximize the average PID of all 10 human-machine pairs for each song and each human transcriber.

4. HYPOTHESES
transcribers. This means that consensus is indeed reflecting the elements of those three transcribers' transcriptions rather than the particular pairs. These results suggest that transcription of pitch contour could be reliable even for non-Western music.

We also analyze some low agreement results. There are 25 pairwise $\kappa$ values lower than 0.4, all of which involved 7 songs (NAIV-033, NAIV-100, NAIV-117, T5431R27, T5482R03, NAIV-015 and NAIV-048). In particular, NAIV-033 (Maya healing song) is a near-monotone chant, and so the degree of agreement by chance is so large that it negates the proportion of agreement. As the unison interval dominates, the note sequences of this song are highly homogeneous (PID > 0.9 for all 6 pairs). Other than this song, the remaining disagreement is mainly caused by disagreement of the pitch rather than segmentation. (cf. supplementary material table S1). In other words, transcribers generally captured the same note events, but the assigned pitch sometimes differs by 1-2 semitones.

Incidentally, we observed that using raw note sequences yielded the median of $\kappa$ very close to zero ($\kappa = -0.019$) due to cases where the tonal center differed by a semitone (and sometimes a whole tone, cf. supplementary material figure S1-S2 for a sample figure and the results). Therefore, our evaluation actually focused on whether relative pitch, or the shape of the pitch contour, matches between transcribers.

5.2 Q2. Which automated method agrees best with transcription of non-Western music by humans?

The right-hand side of Figure 3 shows $\kappa$ and PID obtained by comparing the machine note sequences and our consensus version's note sequences. Contrary to our prediction, there is no evidence for the superiority of TONY except when compared with AD-NNMF and SPICE. The figure also indicates generally low reliability of automated transcription methods (median $\kappa$ values are all below 0.4). In particular, SPICE and AD-NNMF both had median $\kappa$ below 0, suggesting they performed worse than chance. Especially, AD-NNMF failed to pick up notes correctly in many cases and indeed, sometimes the length of note sequence of AD-NNMF is zero (cf. supplementary material figure S8-S9). In such cases, the proportion of agreement between human note sequences also becomes around zero, but chance agreement probability is still positive by its definition, resulting in many negative Kappa values.

In addition, SPICE and CREPE had difficulty estimating F0 of the particular tracks of monophonic singing, which is apparent from the drop in the plot of note sequence length (cf. supplementary material figure S8-S9). As predicted, $\kappa$ of automated methods designed for monophonic vocal melody (i.e. TONY, pYIN, CREPE and SPICE) show a relatively large difference dependent on instrumental accompaniment compared to the other methods.
but STF also suffered from instrumental sounds. (cf. Figure 3 and supplementary material figure S10).

See supplementary material for additional analysis details including results of measuring agreement with Levenshtein distances (which were generally similar to results found using PID).

6. DISCUSSION AND FUTURE WORK

Overall, we observed that the degree of agreement of transcriptions of diverse traditional songs among human transcribers was relatively high (~90% agreement, \( \kappa \approx 0.7 \), Fig. 3), while the degree of agreement between human and automated methods was relatively low (~60% agreement, \( \kappa \approx 0.4 \), Fig. 3). Automated methods where less than 60% of estimated notes agree with human judgments are unlikely to produce satisfactory results for the kinds of tasks we hope to use them for, such as cross-cultural comparison of scale and interval systems [41-42]. Landis and Koch [43] suggested that \( \kappa \) of .61-.8 be considered "substantial" agreement. 21-0.4 as "fair" agreement, but some have suggested that less than .4 is unacceptably low [44]. Our qualitative examination of the transcriptions (e.g., Fig. 2) supports the interpretation that human transcriptions of diverse traditional songs can be sufficiently reliable, but current state-of-the-art automated methods are not. However, high agreement does not necessarily equate to high quality. The quality of transcription may depend on its goal [45-46], so future research should expand on our results to evaluate transcriptions for specific applications (e.g., tonal analysis [41-42]).

Different combinations of human transcribers and songs had varying levels of agreement, but overall the agreement among three female Japanese experts and the consensus transcription by three white American male experts was surprisingly high, with more differences appearing between individuals than between the two groups. Of course, by definition the experts had been trained in Western music and transcription methods - future studies should explore perceptual variability among listeners with varying degrees of training in different musical systems [47].

Disagreement among humans appeared to primarily involve assignment of pitch to different pitch classes. In contrast, disagreement in automated methods appeared to primarily reflect segmentation, rather than F0 estimation. Future studies might be able to clarify this point by collecting both F0 annotations and score transcriptions by humans. This might also allow us to compare our results with more conventional metrics used in research on pitch estimation algorithms such as frame-wise and note-wise F0 agreement and the use of true positive and false positive scores [10] (though we emphasize that our work brings into question the idea that a single 'ground-truth' annotation might even exist that all can agree on for diverse songs). [48] developed a method for evaluating the degree of agreement of F0-estimates among multiple automated methods, and such a method would be especially advantageous to assess the overall reliability of automated methods against global songs once F0 annotation is collected.

We were surprised that all automated methods performed so poorly even for the relatively simple task of transcribing only pitch sequences for monophonic songs. Some might argue it is unfair to evaluate MIR methods designed primarily for F0 transcription of Western instrumental music using symbolic notation transcriptions of (mostly) non-Western songs. Our feelings are somewhat opposite - it is unfair and unethical to limit MIR to a narrow slice of the world's music [49]. Since our goal was to evaluate the ability of existing MIR algorithms to transcribe global songs using symbolic notation, we believe it is fair and necessary to evaluate state of the art algorithms even though - in fact especially because - they were not designed for this application. Our results thus confirm the strong need for automatic music transcription and other MIR tasks to expand algorithms and datasets beyond the traditional focus on Western classical and popular music to be suitable for more diverse musical styles [49]. Moving from a reliance on convenient but restricted datasets (e.g., the MAPS dataset of MIDI-aligned piano recordings commonly used to evaluate automatic transcription [11]) to cross-cultural datasets like the ones presented here and elsewhere [9, 16, 50] will be essential for diversifying MIR.

The formalization of a general algorithm that agrees with human pitch recognition and note segmentation is an ongoing challenge related to a central issue in MIR: the "correctness" of the algorithm depends on the degree of perceptual variability in the human ground-truth data [51]. Thus, accounting for diversity and subjectivity in human transcriptions is equally critical to advance research on the automatic analysis of music. For example, while we found relatively high agreement among expert transcribers using Western 12-TET notation, we do not know whether the singers whose songs we transcribed would agree with our transcriptions, or whether transcription using a different notation system (e.g., Middle Eastern 24-note microtonal notation, ‘Are’Are 7-note equiheptatonic notation [52], Killick’s “global notation” [53], etc.) would give better or worse results. We see our current results using 12-TET - with all its known problems and cultural baggage [1-5, 45-46] - as a baseline against which future studies can test whether other methods of cross-cultural transcription may be able to improve.

Furthermore, here we solely focused on pitch, but a more comprehensive description of music necessitates other dimensions such as rhythm, timbre, and social context [54]. Other cross-cultural systems of music analysis such as Cantometrics [54-55] and CantoCore [56] have been designed to capture such features. Somewhat counterintuitively, our current results show substantially higher agreement using Western staff notation to analyze a global song sample (\( \kappa \approx 0.7 \)) than was found using these cross-cultural song classification systems (\( \kappa \approx 0.3-0.5 \) [8, 16, 56]). This suggests a need for MIR to better account for diversity in human ground-truth representations of all dimensions of music, not only pitch [57].

Musical diversity is a crucial challenge and opportunity for MIR. Quantifying diversity in human “ground-truth” cross-cultural data is an important first step for diversifying MIR. Our study demonstrates that there is still substantial room for improvement for automated methods of music transcription, and provides quantitative estimates of diversity among human transcriptions to help guide development of future MIR methods.
7. AUTHOR CONTRIBUTIONS

8. ACKNOWLEDGEMENTS
We thank Olga Velichkina and members of the Keio SFC CompMusic and NeuroMusic labs for feedback on earlier versions of this manuscript. This work was supported by Grant-in-Aid no. 19KK0064 from the Japan Society for the Promotion of Science and by startup grants from Keio University (Keio Global Research Institute, Keio Research Institute at SFC, and Keio Gijuku Academic Development Fund) to P.E.S.

9. DATA/CODE AVAILABILITY
Audio files, transcriptions (individual and consensus), aligned note sequences, and analysis scripts are available at https://github.com/comp-music-lab/agreement-human-automated (codes, transcriptions, sequence data), https://doi.org/10.5281/zenodo.4941863 (audio). The PDF of the supplementary material is also available at the above GitHub link.

10. REFERENCES


