

# Reliability of automated and human transcriptions of non-Western music

Yuto Ozaki,<sup>1</sup> John McBride,<sup>2</sup> Emmanouil Benetos,<sup>3</sup> Peter Pfordresher,<sup>4</sup> Joren Six,<sup>5</sup> Adam T. Tierney,<sup>6</sup> Polina Proutskova,<sup>7</sup> Emi Sakai,<sup>8</sup> Haruka Kondo,<sup>9</sup> Haruno Fukatsu,<sup>10</sup> Shinya Fujii,<sup>11</sup> Patrick E. Savage<sup>12</sup>

<sup>1,9,10</sup>Graduate School of Media and Governance, Keio University, Japan

<sup>2</sup>Center for Soft and Living Matter, Institute for Basic Science, South Korea

<sup>3,7</sup>School of Electronic Engineering and Computer Science, Queen Mary University of London, UK

<sup>4</sup>Department of Psychology, University at Buffalo, USA

<sup>5</sup>Department of Art, Music and Theatre Sciences, Ghent University, Belgium

<sup>6</sup>Department of Psychological Sciences, Birkbeck, University of London, UK

<sup>8</sup>[No affiliation], Japan

<sup>11,12</sup>Faculty of Environment and Information Studies, Keio University, Japan

<sup>1</sup>yozaki@sfc.keio.ac.jp, <sup>2</sup>jmmcbride@protonmail.com, <sup>3</sup>emmanouil.benetos@qmul.ac.uk, <sup>4</sup>pqp@buffalo.edu, <sup>5</sup>joren.six@ugent.be,

<sup>6</sup>a.tierney@bbk.ac.uk, <sup>7</sup>polinaproutskova@gmail.com, <sup>8</sup>emymelody@gmail.com, <sup>9</sup>haruka\_kondo@keio.jp, <sup>10</sup>haruo.violin@keio.jp,

<sup>11</sup>sfujii@sfc.keio.ac.jp, <sup>12</sup>psavage@sfc.keio.ac.jp

## Background

Cross-cultural analysis is essential to explore diversity and universality of music perception and cognition. 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. Automated methods of music transcription might potentially solve these problems.

## Aims

The main objective is to evaluate the degree of agreement among human and automated transcriptions for a global song sample. To do so, we evaluated the degree of agreement among transcriptions created by human transcribers and the agreement among human transcribers and 10 automatic music transcription/vocal melody extraction algorithms. In particular, we created consensus transcriptions by merging the 3 individual transcriptions and compared the consensus transcriptions and the transcriptions contained in the Natural History of Song (Mehr et al., 2019) dataset for human-human agreement evaluation. Similarly, we compared the consensus transcriptions and machine transcriptions for the human-machine cases.

## Method

Twelve-tone equal temperament (12-TET) with A4 = 440 Hz is used to transcribe audio into staff notation for humans. 12-TET is also applied to automated methods to standardize their outputs along with various post-processing procedures. While binning continuous F0 into a discrete set of 12 100-cent intervals does not always correspond to the actual scales/modes, it helps to obtain an interpretable symbolic representation of melody and enables us to compare our results with Mehr et al (2019)'s transcriptions. This study focuses on the degree of agreement of the sequence of pitched notes. Hence, we discard temporal/rhythmic information and we only extract pitch class from transcriptions to create a sequence of notes. Agreement among human and machine note sequences is evaluated by Fleiss' Kappa and percent identity (PID). We use the Needleman-Wunsch algorithm to align note sequences. We sampled excerpts of 32 audio recordings of traditional solo singing from Global Jukebox (Wood et al., 2021) and Natural History of Song (Mehr et al., 2019) from eight regions of the world (17 a cappella, 15 accompanied).

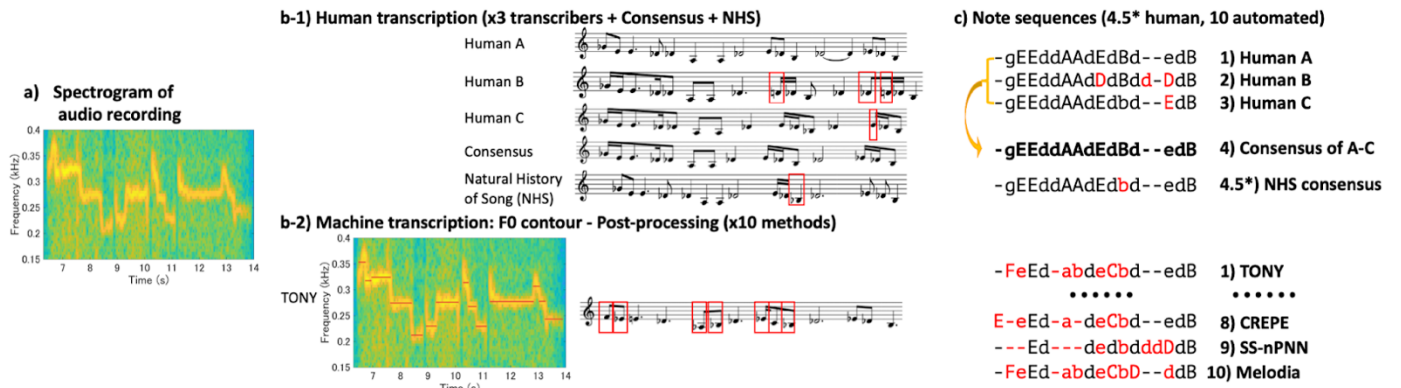


Figure 1. Schematic overview of the evaluation.

## Results

We observed that the degree of agreement of transcriptions of diverse traditional songs among human transcribers was relatively high (median  $\kappa \sim 0.7$ , PID  $\sim 90\%$ ), while the degree of agreement between human and automated methods was relatively low

(median  $\kappa < 0.4$ , PID  $< 60\%$ ). 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.

## Conclusions and Implications

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 the research on automated methods for music analysis: the “correctness” of the algorithm depends on the degree of perceptual variability in the human ground-truth data (Flexer & Grill, 2016). Thus, future studies should explore perceptual variability among listeners with varying degrees of training in different musical systems. 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 would give better or worse results. Furthermore, here we solely focused on pitch, but a more comprehensive analysis of music perception and cognition necessitates other dimensions such as rhythm, timbre, and social context. Overall, establishing objective, reliable methods for interpretable cross-cultural music representation is critical for creating a more diverse, cross-culturally connected science of music cognition.

## References

- 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. doi:10.1080/09298215.2016.1200631
- Mehr, S. A., Singh, M., Knox, D., Ketter, D. M., Pickens-Jones, D., Atwood, S., ..., Glowacki, L. (2019). Universality and diversity in human song. *Science*, 366, eaax0868. doi:10.1126/science.aax0868
- Wood, A. L. C., Kirby, K. R., Ember, C. R., Silbert, S., Daikoku, H., Paulay, F., ..., Savage, P. E. (2021). The Global Jukebox: A public database of performing arts and culture. *PsyArXiv Preprint*. doi:org/10.31234/osf.io/4z97j

**Keywords:** Pitch, Melody, Music information retrieval, Automatic music transcription, Cross-cultural analysis