AGREEMENT AMONG HUMAN AND AUTOMATED TRANSCRIPTIONS OF GLOBAL SONGS

Yuto Ozaki[1], John McBride[2], Emmanouil Benetos[3], Peter Q. Pfordresher[4], Joren Six[5], Adam T. Tierney[6], Polina Proutskova[3], Emi Sakai[7], Haruka Kondo[1], Haruno Fukatsu[1], Shinya Fujii[1], Patrick E. Savage[1*]

[1]Keio University, Fujisawa, Japan
[2]Center for Soft and Living Matter, Institute for Basic Science, South Korea
[3]Queen Mary University of London, UK
[4]University at Buffalo, NY, USA
[5]Ghent University, Belgium
[6]Birkbeck, University of London, UK
[7]No affiliation
[*]psavage@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 ("Rumanian 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.

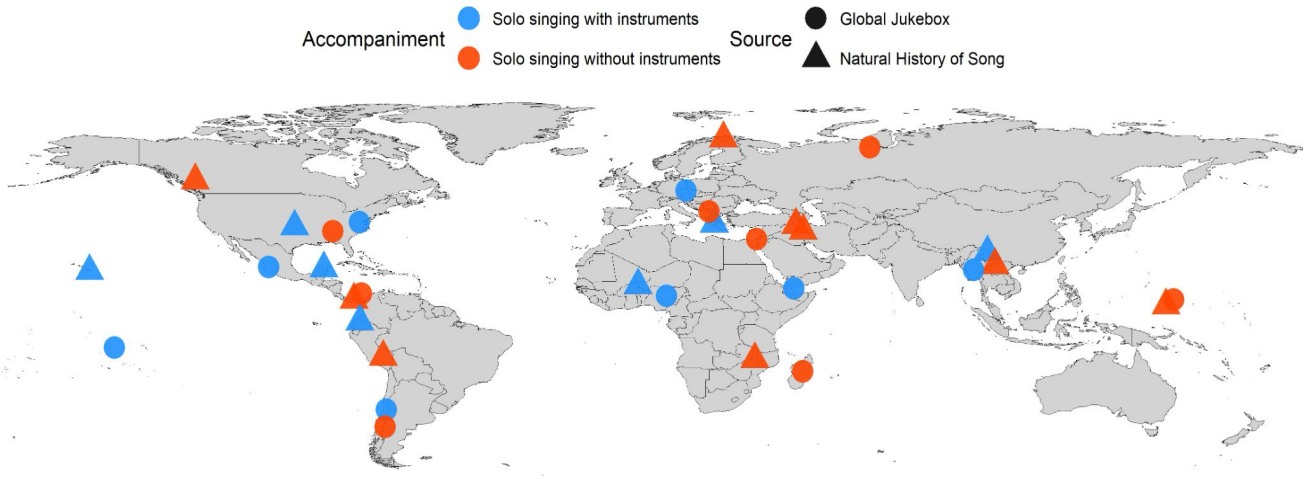

Figure 1. Map of the 32 songs transcribed and analyzed.

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 in-strumental 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

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].

Method	Target sound	Unit	Category
pYIN [17]	Monophonic vocal	Frame	Non data-driven: parameters specified manually
TONY [18]	Monophonic vocal	Note	
Melodia [22]	Vocal melody	Frame	
STF [23]	Multiple 12-tone ET	Frame	
CREPE [24]	Monophonic vocal	Frame	Data-driven: parameters optimized by training with datasets.
SPICE [25]	Monophonic vocal	Frame	
SS-nPNN [26]	Vocal melody	Frame	
AD-NNMF [27]	Multiple piano sound	Note	
OAF [28]	Multiple piano sound	Note	
madmom [29]	Multiple piano sound	Note	

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 Fleiss' 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/jh7t5/

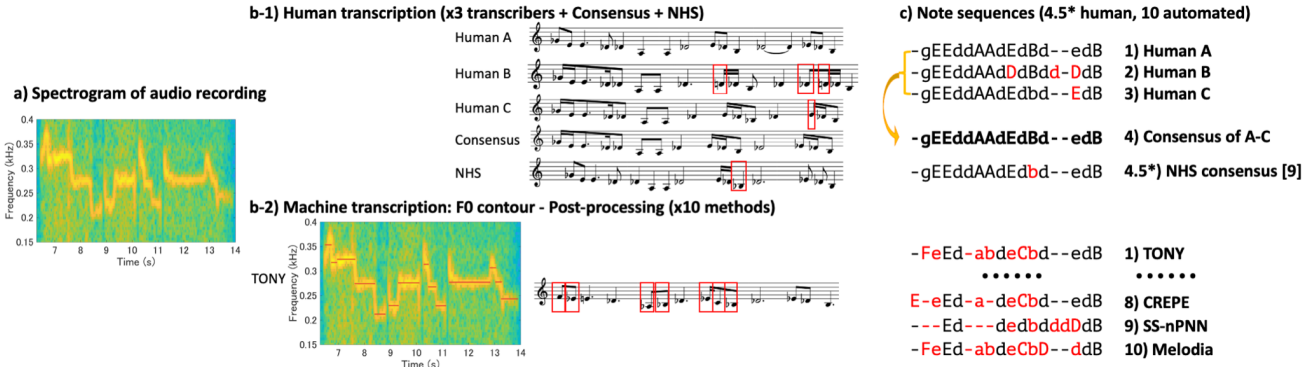

Figure 2. Overview of the agreement evaluation using an example 8-second excerpt from NAIV-075 (Healing song, Kwakwaka'wakw people, 00:06-00:14 from https://osf.io/y29wp). Red indicates disagreement with our new consensus transcription (#4, made by combining the three individual transcriptions #1-3). 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, #4.5 (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.

pitch (i.e., substitution) or note segmentation (i.e., insertion/deletion). We also report such quantities by using percent identity (PID) [36-37] (cf. Fig. 2 for an example and for Supplementary Material for detailed explanation and additional analyses using Levenshtein distances). Although our approach did not utilize the real time information of note events, we confirmed it can still make meaningful alignment of notes as in the previous study [37] from the pilot experiment. Meanwhile, we also admit there is the technical difficulty of the extraction of note event timing (i.e. segmentation of sounds). Importantly, our work differs from the related studies evaluating the agreement between human and automated methods' melody annotation [38-39] in aiming symbolic-level note comparison rather than frame-level F0 comparison. In addition, previous studies only reported individual metrics (e.g. either distance-like metrics or inter-rater reliability, but not both), while our study explored agreement of symbolic-level melody using multiple metrics.

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 κ, 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

We pre-registered[2] the following two primary hypotheses and 10 corresponding predictions based on pilot analysis of 4 songs not included in our main analyses:

H1: Human transcriptions are sufficiently reliable. This predicts a Fleiss' Kappa coefficient significantly greater than 0 when comparing our consensus transcription against the consensus transcriptions of Mehr et al. [9]. Note sequences including unison intervals are used.

H2: TONY is the most reliable method of automated singing transcription. We predicted this because unlike other methods TONY was designed to perform note segmentation for human vocal melody, better matching human standards for transcription. This predicts that Fleiss' Kappa comparing TONY with our consensus note sequences will be significantly greater than for the other 9 algorithms when evaluated against the note sequences including unison intervals.

5. RESULTS

5.1 Q1. To what degree do humans' transcriptions agree?

The left-hand side of Figure 3 shows inter-rater reliability and percent identity results comparing human transcribers. As predicted, there was significant agreement between our new consensus transcriptions and the pre-existing NHS consensus transcriptions (median $\kappa = .74$, $p < .001$; median PID = 88%). When we compare the results using individual transcriptions rather than consensus transcriptions, we see that agreement is slightly lower but still relatively high (lowest median κ of .64 and PID 83% for Transcribers A & B). The left-most two boxes show that individual vs. consensus yields higher agreement than individual vs. individual combinations (e.g. A-Cons, A-B, A-C) for all

[2] https://osf.io/bjemd

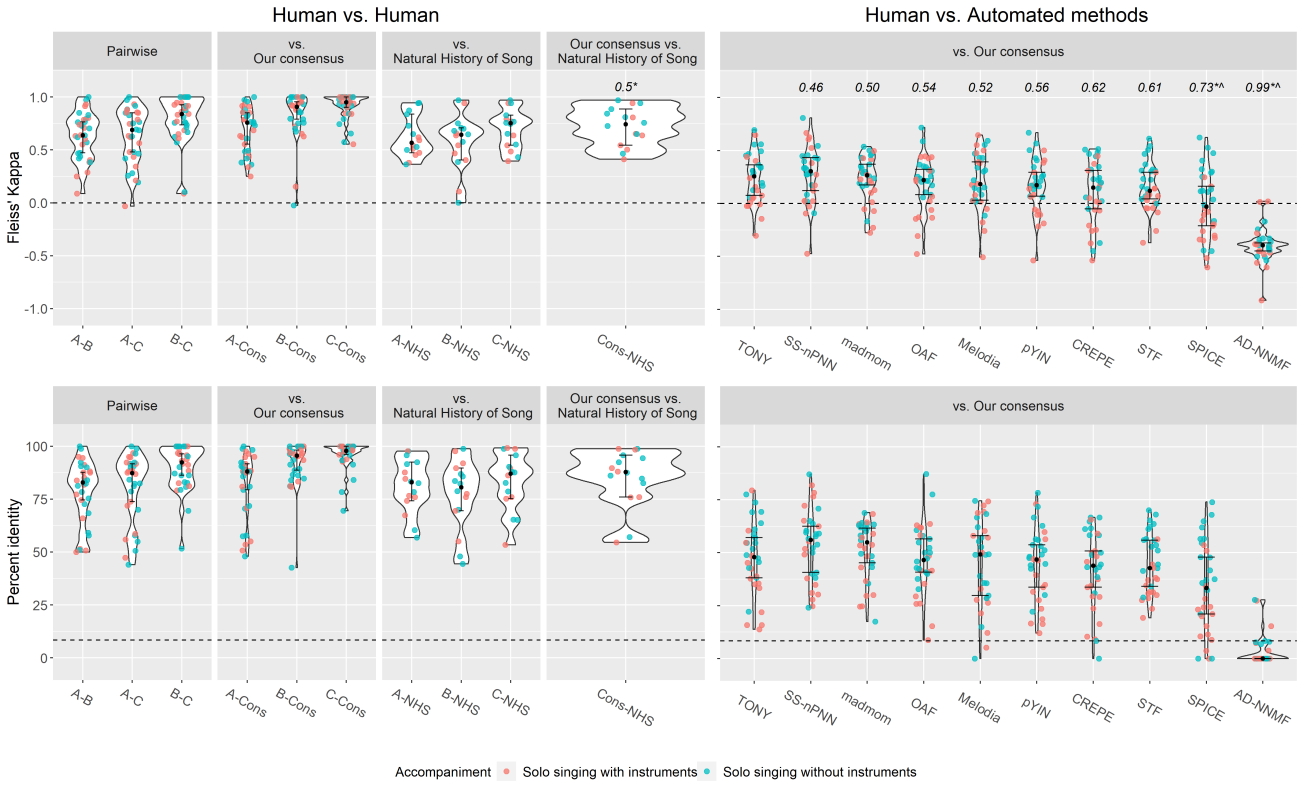

Figure 3. Agreement among human-human and human-machine transcribed note sequences. "A", "B", and "C" represent the three individual transcribers. The dashed line at κ = 0 and 8.3% identity indicates chance levels of agreement. The numbers appearing above the violin plots indicate effect sizes for our 10 pre-registered predictions, and * and ^ indicate significant p-value and posterior probability, respectively (cf. text for details). Black circles indicate medians, and bars represent 95% confidence intervals of the median [40]. Results using alternative transcription methods, and full p-value and posterior probabilities are available in the supplementary material (figure S3-S7 and table S2-S3).

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 κ 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 κ very close to zero (κ = -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 κ 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 κ values are all below 0.4). In particular, SPICE and AD-NNMF both had median κ 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, κ 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, κ ~0.7; Fig. 3), while the degree of agreement between human and automated methods was relatively low (< 60% agreement, κ <.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 κ 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 one 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 (κ ~0.7) than was found using these cross-cultural song classification systems (κ ~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

Conceptualization: P.E.S, J.M, P.Q.P., J.S., A.T.T., S.F., E.B., Y.O., P.P.; Transcription: E.S., H.K., H.F.; Methodology / Analysis / Investigation / Visualization: Y.O., P.E.S., E.B., J.M.; Project administration / Supervision / Funding acquisition: P.E.S.; Writing – original draft: Y.O., P.E.S.; Writing – review & editing: J.M., E.B., P.Q.P., J.S., A.T.T., P.P., S.F.

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

[1] B. Nettl, *The Study of Ethnomusicology: Thirty-Three Discussions, 3rd ed.*, Champaign, IL, USA: University of Illinois Press, 2015.

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

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

[4] 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.

[5] E. Benetos, S. Dixon, Z. Duan, and S. Ewert, “Automatic music transcription: An overview,” *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 20–30, 2019.

[6] M. Müller, E. Gómez, and Y. Yang, Eds., "Computational methods for melody and voice processing in music recordings (Dagstuhl seminar 19052)," *Dagstuhl Reports*, vol. 9, no. 1, pp. 125–177, 2019.

[7] Z. Rafii, A. Liutkus, F. Stöter, S. I. Mimilakis, D. FitzGerald, and B. Pardo, "An overview of lead and accompaniment separation in music," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 8, pp. 1307-1335, 2018, doi: 10.1109/TASLP.2018.2825440.

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

[9] S. A. Mehr et al., "Universality and diversity in human song," *Science*, vol. 366, eaax0868, 2019, doi: 10.1126/science.aax0868.

[10] A. Ycart, L. Liu, E. Benetos, and M. T. Pearce, "Investigating the perceptual validity of evaluation metrics for automatic piano music transcription," *Transactions of the International Society for Music Information Retrieval*, vol. 3, no. 1, pp. 68–81, 2020, doi: 10.5334/tismir.57.

[11] V. Emiya, R. Badeau, and B. David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1643–1654, 2010.

[12] A. Holzapfel, and E. Benetos, "Automatic music transcription and ethnomusicology: A user study," in *Proc. 20th International Society for Music Information Retrieval Conference*, Delft, The Netherlands, 2019, pp. 678-684.

[13] E. Gómez, and J. Bonada, "Towards computer-assisted flamenco transcription: An experimental comparison of automatic transcription algorithms as applied to a cappella singing," *Computer Music Journal*, vol. 37, no. 2, pp. 73–90, 2013.

[14] S. Cottrell, "Big music data, musicology, and the study of recorded music: Three case studies," *The Musical Quarterly*, vol. 101, no. 2-3, pp. 216-243, doi: 10.1093/musqtl/gdy013.

[15] L. Tilley, "Analytical ethnomusicology: How we got out of analysis and how to get back in," in *Springer Handbook of Systematic Musicology*, R. Bader, Eds. Berlin, Germany: Springer, 2018, pp. 953–977.

[16] A. L. C. Wood *et al.*, “The Global Jukebox: A public database of performing arts and culture,” *PsyArXiv Preprint.* doi: org/10.31234/osf.io/4z97j.

[17] M. Mauch, and S. Dixon, "PYIN: A fundamental frequency estimator using probabilistic threshold distributions," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Florence, Italy, 2014, pp. 659–663

[18] M. Mauch et al., “Computer-aided melody note transcription using the tony software: Accuracy and efficiency,” in *Proc. 1st International Conference on Technologies for Music Notation and Representation*, Paris, France, 2015.

[19] R. Nishikimi, E. Nakamura, M. Goto, and K. Yoshii, "Audio-to-score singing transcription based on a CRNN-HSMM hybrid model," *APSIPA Transactions on Signal and Information Processing*, vol. 10, no. e7, pp. 1-13, 2021, doi: 10.1017/ATSIP.2021.4.

[20] R. Gunter, C. Carvalho, and P. Smaragdis, "Towards end-to-end polyphonic music transcription: Transforming music audio directly to a score," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New York, USA, 2017, pp. 151-155, doi: 10.1109/WASPAA.2017.8170013.

[21] M. A. Román, A. Pertusa, and J. Calvo-Zaragoza, "A holistic approach to polyphonic music transcription with neural networks," in *Proc. 20th International Society for Music Information Retrieval Conference*, Delft, The Netherlands, 2019, pp. 731-737.

[22] J. Salamon, and E. Gomez, "Melody extraction from polyphonic music signals using pitch contour characteristics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1759-1770, 2012, doi: 10.1109/TASL.2012.2188515.

[23] L. Su, and Y. Yang, "Combining spectral and temporal representations for multipitch estimation of polyphonic music," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 10, pp. 1600-1612, 2015, doi: 10.1109/TASLP.2015.2442411.

[24] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "Crepe: A convolutional representation for pitch estimation," in *Proc. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, Calgary, AB, Canada, 2018, pp. 161-165, doi: 10.1109/ICASSP.2018.8461329.

[25] B. Gfeller, C. Frank, D. Roblek, M. Sharifi, M. Tagliasacchi, and M. Velimirović, "SPICE: Self-supervised pitch estimation," *IEEE/ACM Transactions on Audio, Speech, and Language*, vol. 28, pp. 1118-1128, 2020, doi: 10.1109/TASLP.2020.2982285.

[26] W.-T. Lu, and L. Su, "Vocal melody extraction with semantic segmentation and audio-symbolic domain transfer learning," in *Proc. 19th International Society for Music Information Retrieval Conference*, Paris, France, 2018, pp. 521–528.

[27] T. Cheng, M. Mauch, E. Benetos, and S. Dixon, "An attack/decay model for piano transcription," in *Proc. 17th International Society for Music Information Retrieval Conference*, New York, NY, USA, 2016, pp. 584–590.

[28] C. Hawthorne et al., "Onsets and Frames: Dual-objective piano transcription," in *Proc. 19th International Society for Music Information Retrieval Conference*, Paris, France, 2018, pp. 50-57.

[29] S. Bock, F. Korzeniowski, J. Schluter, F. Krebs, and G. Widmer, "Madmom: A new Python audio and music signal processing library," in *Proc. 24th ACM International Conference on Multimedia*, Amsterdam, Netherlands, 2016.

[30] R. Ambrazevičius: "The perception and transcription of the scale reconsidered: Several Lithuanian cases," *The World of Music*, vol. 47, no. 2, pp. 31-53, 2005.

[31] P. Q. Pfordresher, S. Brown, K. M. Meier, M. Belyk, and M. Liotti, "Imprecise singing is widespread," *The Journal of the Acoustical Society of America*, vol. 128, no. 4, pp. 2182-2190, 2010.

[32] P. Larrouy-Maestri, Y. Lévêque, D. Schön, A. Giovanni, and D. Morsomme, "The evaluation of singing voice accuracy: A comparison between subjective and objective methods," *Journal of Voice*, vol, 27, no. 2, pp. 259.e1-259.25, 2013, doi: 10.1016/j.jvoice.2012.11.003.

[33] MuseScore: https://musescore.org/ja (accessed Feb. 25, 2021).

[34] S. B. Needleman, and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443–453, 1970.

[35] K. L. Gwet, *Handbook of Inter-Rater Reliability. The Definitive Guide to Measuring the Extent of Agreement Among Raters, 4th edition*. Gaithersburg, MD, USA: Advanced Analytics, LLC., 2015.

[36] A. C. W. May, "Percent sequence identity: The need to be explicit," *Structure*, vol. 12, no. 5, pp. 737–738, 2004, doi: 10.1016/j.str.2004.04.001.

[37] P. E. Savage, and Q. D. Atkinson: "Automatic tune family identification by musical sequence alignment," in *Proc. 16th International Society for Music Information Retrieval Conference*, Málaga, Spain, 2015, pp. 162-168.

[38] J. J. Bosch, and E. Gómez, "Melody extraction in symphonic classical music: a comparative study of mutual agreement between humans and algorithms," in *Proc. 9th Conference on Interdisciplinary Musicology*, Berlin, Germany, 2014.

[39] S. Balke, J. Abeßer, J. Driedger, C. Dittmar, and M. Müller, "Towards evaluating multiple predominant melody annotations in Jazz recordings," in *Proc. 17th International Society for Music Information Retrieval Conference*, New York City, NY, USA, 2016, pp. 246–252.

[40] M. J. Campbell, and M, J. Gardner, "Calculating confidence intervals for some non-parametric analyses," *British Medical Journal*, vol. 296, no. 6634, pp. 1454-1456, doi: 10.1136/bmj.296.6634.1454.

[41] J. M. McBride and T. Tlusty, "Cross-cultural data suggests musical scales evolved to maximise imperfect fifths," *arXiv Preprint*, 2020. arXiv:1906.06171.

[42] J. Kuroyanagi *et al.*, "Automatic comparison of human music, speech, and bird song suggests uniqueness of human scales," in *Proc. 9th International Workshop on Folk Music Analysis (FMA2019)*, Birmingham, UK, 2019, pp. 35–40.

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

[44] 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.

[45] C. Seeger, “Prescriptive and descriptive music-writing,” *Music. Q.*, vol. 44, no. 2, pp. 184–195, 1958.

[46] M. Hood, *The Ethnomusicologist*. New York: McGraw-Hill, 1971.

[47] N. Jacoby et al., “Universality and cross-cultural variation in mental representations of music revealed by global comparison of rhythm priors,” *PsyArXiv Preprint, 2021.* doi: 10.31234/osf.io/b879v.

[48] S. Rosenzweig, F. Scherbaum, and M. Müller, "Reliability assessment of singing voice F0-estimates using multiple algorithms," in *Proc. 2021 IEEE International Conference on Acoustics, Speech and Signal Processing*, Ontario, Canada, 2021, pp. 261-265.

[49] G. Born, “Diversifying MIR: Knowledge and real-world challenges, and new interdisciplinary futures,” *Trans. Int. Soc. Music Inf. Retr.*, vol. 3, no. 1, pp. 193–204, 2020.

[50] 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. Oxford University Press, in press. doi: 10.31235/osf.io/nxtbg.

[51] 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, doi: 10.1080/09298215.2016.1200631.

[52] H. Zemp, and V. Malkus, “Aspects of ’Are ’Are musical theory,” *Ethnomusicology*, vol. 23, no. 1, pp. 5–48, 1979.

[53] A. Killick, “Global notation as a tool for cross-cultural and comparative music analysis,” *Analytical Approaches to World Music*, vol. 8, no. 2, pp. 235–279, 2021.

[54] P. E. Savage, “Alan Lomax’s Cantometrics Project: A comprehensive review,” *Music & Science*, vol. 1, pp. 1–19, 2018.

[55] A. Lomax, and V. Grauer, “The Cantometric coding book,” in *Folk Song Style and Culture*, A. Lomax, Ed. Washington, DC, USA: American Association for the Advancement of Science, 1968, pp. 34–74.

[56] P. E. Savage, E. Merritt, T. Rzeszutek, and S. Brown, “CantoCore: A new cross-cultural song classification scheme,” *Analytical Approaches to World Music*, vol. 2, no. 1, pp. 87–137, 2012.

[57] H. Daikoku, S. Ding, U. S. Sanne, E. Benetos, A. L. Wood, S. Fujii, and P. E. Savage: “Human and automated judgements of musical similarity in a global sample,” *PsyArXiv Preprint*, 2020*,* doi: https://doi.org/10.31234/osf.io/76fmq.

Supplementary Materials for

AGREEMENT AMONG HUMAN AND AUTOMATED TRANSCRIPTIONS OF GLOBAL SONGS

A. PROCEDURE OF CREATING CONSENSUS NOTE SEQUENCE

We create a consensus transcription of each song by the following steps. Firstly, we automatically align the note sequences and then perform manual correction with rhythmic information. Secondly, disagreements of each note among note sequences are resolved by majority rule, following [9]. If there is a note that is different in all three note sequences, we ask our collaborators via email to choose which note would fit the consensus notes selected by the majority rule. If disagreement still remains, we choose the median of the pitch from the disagreeing notes. However, subjective decisions were made when disagreement involved alignment gaps. We then asked transcribers by email to confirm the soundness of the resultant consensus transcriptions, and further updated some transcriptions based on this feedback.

B. DETAILS OF POST-PROCESSING PROCEDURE

In order to standardize the output of each method, we applied the following processes.

1. For methods that do not quantize F0 to twelve-tone equal temperament, the estimated F0 is rounded to the nearest frequency of the twelve-tone equal temperament.
2. For methods that do not estimate note duration or note tracking, a median filter with the length of 0.25 seconds is applied to smooth the pitch contour. Furthermore, the sequences of F0 shorter than 0.15 seconds are ignored from transcription. 0.15 is determined to make the length of note sequence similar to the humans' sequences. These parameters were tuned to minimize the possibility that the automated methods would produce long sequences made up of unrealistically short notes as a by-product of the instability of pitch targets in human singing. If the unit of the discrete time interval of generated time-frequency representation is less than 0.01 second, decimation is applied to make the interval close to 0.01 second to smooth the pitch contour.
3. For methods predicting multiple pitches in a single timeframe, we apply the following steps to obtain the stream of single pitch prediction. Firstly, we observe that these methods tend to predict an overtone as a separate note, so the frequency range of the melody is manually specified, and the F0 prediction out of this range was removed. After that, the Viterbi algorithm is applied to the remaining multi-pitch F0 prediction results to obtain the dominant time-frequency energy sequence as a melody [58].
4. Regarding CREPE, F0s having a confidence score larger than or equal to 0.8 are picked up. Note that there is no guideline of what value to be used for a threshold. If we used a lower threshold, the final note sequence would become longer due to including more pitches, and that would result in lower PID and Kappa than the currently presented results.
5. We use a song excerpt as the input of automated methods to obtain the pitch estimation of a specified 14-second segment. However, pitch estimation process would depend on the information available on the broader time range of audio data to estimate the F0 of local time-frame, so feeding an entire song as input and extracting the target segment from its output will produce different pitch estimation results. In this study, we only have the excerpt of songs regarding the NHS, so we decided to consolidate the input by an audio excerpt.

Incidentally, OAF estimates onset and offset of note, but it is fairly precise, so the above post-processing is applied to make a more meaningful comparison with the other methods.

C. SEQUENCE ALIGNMENT METHOD

We perform pairwise alignment to create the alignment of note sequences by the Needleman-Wunsch algorithm using 0.0 for gap opening penalty, -1.0 for gap extension penalty and -1.0 for mismatch (substitution) penalty. This is a linear gap setting, and we choose this setting that makes the alignment score equivalent to Levenshtein distance whose operations (i.e. insertion, deletion, substitution) are all equally weighted. We use octave information for the evaluation, so the element of the sequence consists of two characters: pitch class and octave level (e.g. "A4"). When multiple sequence alignment is necessary for creating the baseline of consensus note sequences, we use the center star method to solve alignment heuristically since the computation of the global optimal multiple sequence alignment is not feasible due to its computational complexity [59-60]. The center sequence is determined by the sum-of-pairs scoring [59-60], and each score is calculated by the Needleman-Wunsch algorithm as described above.

D. METRICS OF AGREEMENT AMONG SEQUENCES

Regarding the computation of Fleiss' Kappa, we regard note transcriptions as a transcriber's categorization of the F0 of a given note. We do not apply partial agreement in this study. The length of the sequence corresponds to the number of subjects, and the number of sequences corresponds to the number of raters. When calculating the inter-rater reliability coefficients, we also treat gaps inserted during the alignment as coding rather than absence. Gap insertion indicates that some transcribers interpret the sound as a pitch, but the others do not, which we treat as a coding disagreement.

On the other hand, the practice of reporting the raw percent agreement score along with inter-rater reliability coefficients is also discussed due to its simplicity [35, 61]. Percent identity (PID) measures the proportion of concordance elements of two sequences which is conceptually equivalent to percent agreement, and we use this metric to evaluate how much two note sequences are identical. In the case of multiple sequences appearing in group-wise agreement evaluation, we average the PID by all combinations of pairs in the sequences. PID has originally been used in the computational analysis of protein and DNA sequences to express the similarity between two sequences [36, 60, 62], as well as the comparative study of traditional music melodies [37]. There are several variations in PID [36], and we use the following version.

$$\begin{aligned} \mathrm{PID} &= 100\left\{\frac{N_{\mathrm{ID}}}{0.5(L_1+L_2)}\right\} \\ N_{\mathrm{ID}} &:= \text{Number of identical notes} \\ L &:= \text{Length of sequence} \end{aligned} \tag{1}$$

Although Kappa coefficients and PID can provide the reliability of agreement and the proportion of equality of note sequences respectively, these quantities do not tell how many notes actually differ between note sequences. Therefore, we also use Levenshtein distance to quantify such difference by the number of insert/delete/substitution operations. The penalty of each operation is equally weighted by 1. The score is also averaged in groupwise evaluation cases as well as PID.

E. **STATISTICAL ASSUMPTIONS OF THE TESTS**

Inter-rater reliability coefficients, PID, and Levenshtein distance all quantify the degree of concordance among sequences. The underlying distribution of inter-rater reliability coefficients is considered to depend on the raters (i.e. transcribers) and subjects (audio recording) [35]. Furthermore, our agreement metrics are collected from various combinations of transcribers and audio samples, and the domain of Kappa is finitely bounded, so the resultant distributions of agreement metrics would not necessarily fit normal or location-scale family distributions.

Based on the above assumption, we consider the appropriate testing methods to handle the metrics to be nonparametric methods. We choose the sign test for one-sample test case and the two-sample Anderson-Darling test [63] and two-sample Bayesian nonparametric testing using Pólya trees [64] for two-sample test scenarios. Regarding the two-sample test, we assess the probability of type I error by the two-sample Anderson-Darling test. Besides, to complement the lack of information about how much we can be confident in accepting alternative hypotheses, we also employ Bayesian hypothesis testing. Although these two tests are different procedures, both are proved to be asymptotically consistent under the null hypothesis ($F(x) = G(x)$) and the alternative hypothesis ($F(x) \neq G(x)$) [63-64]. Please refer to the next section for the detailed setting of Bayesian nonparametric testing.

Regarding the effect size to be used for our nonparametric tests, we choose the departure from the expected proportion under the null hypothesis proposed by Cohen [65] for the one-sample test and the probability-based effect size measure A which is known as the probability of one group's superiority over another for two-sample tests [66]. The departure effect size (or Cohen's g) in our study can be interpreted as follows. The sign test uses the number of samples whose value is larger than the expected median under the null hypothesis as test statistics. If the null hypothesis of the sign test is true, then the proportion of data (i.e. κ in our case) above the expected median (i.e. 0 in our case) should be around 50% of all samples. However, if the actual median is larger than 0, then the proportion of samples above the expected median would be larger than 0.5. We calculate the proportion of samples larger than 0 and show the difference between that proportion and the expected proportion under the null hypothesis (i.e. 0.5). Note that in this case, the range of the effect size is from 0 to 0.5 and Cohen [67] suggests interpreting the value larger than 0.25 as the existence of a "Large" effect.

The probability-based effect size uses empirical distributions of data to quantify how much data in a group takes a larger value than another group, and it is robust to violations of the parametric assumptions. We use this effect size to interpret how much TONY's κ is large compared to the others. Note that A can be converted to a common standardized mean difference such as Cohen's d if the normality assumption of data holds [66].

In summary, we put non-normality assumptions for the distributions of κ. Thus, we chose testing methods including Bayesian tests and effect size from nonparametric techniques. We performed the one-tailed one-sample sign test assuming the median of Fleiss' Kappa to be 0 as a null hypothesis for the hypothesis testing of human-human agreement evaluation. Regarding the hypothesis testing of examining the automated method producing transcriptions that best agree with humans' transcriptions, the null hypothesis to be tested is $F_{\mathrm{TONY}}(\kappa) = G_{\mathrm{OTHER}}(\kappa)$, which is the 9 two-sample tests of comparing the empirical distribution of κ by TONY and the others. The superiority of TONY can be quantified by whether the probability-based effect size measure A exceeds 0.5 or not.

F. SETTING OF BAYESIAN NONPARAMETRIC TESTING

We set c = 1and the normal distribution as the centering distribution as the parameters of the Pólya trees (see [64] for the definition of parameterization of this test). However, we use the mean and standard deviation to create partitions of samples instead of the median and quantiles used in the original study. We set the equal probability for the null hypothesis and the alternative hypothesis (= 0.5) as the prior distribution of our Bayesian hypothesis testing, so the posterior odds are equal to Bayes factor.

G. CONTROL OF SIMULTANEOUS INFERENCE

There are 10 null hypothesis significance tests in our analysis: one-sample Sign test × 1 + two-sample Anderson-Darling test × 9 (machine pairs). Since our discussion on the reliability of transcription is interrelated to these test results, we use the False Discovery Rate method to control the p-value threshold for all hypothesis tests regarding these as multiple testing and simultaneous inference. In particular, we will use the Benjamini–Hochberg step-up procedure [68] at level $\alpha = 0.05$ as the threshold to determine the rejection of 10 null hypothesis testing. Incidentally, we will interpret that the Bayesian test at least substantially supports the alternative hypothesis if the posterior probability exceeds 0.8 which corresponds to the Bayes factor = 4 in our setting (i.e. the prior distribution being equally weighed to the null and alternative hypothesis).

H. SEMITONE DISCREPANCY

```
BeeeeBdBBeeBBeeBBBeeeeBdBBeeBBeeBBeeeeeBdBBe-
bDDDDbCbbDDbbDDbbbDDDDbCbbDDbbDDbbbDDDDbCbbGC
```

Figure S1. Example of semitone discrepancy (NAIV-100). Octave information is omitted for visibility.

I. RESULTS OF AGREEMENT USING RAW NOTE SEQUENCES

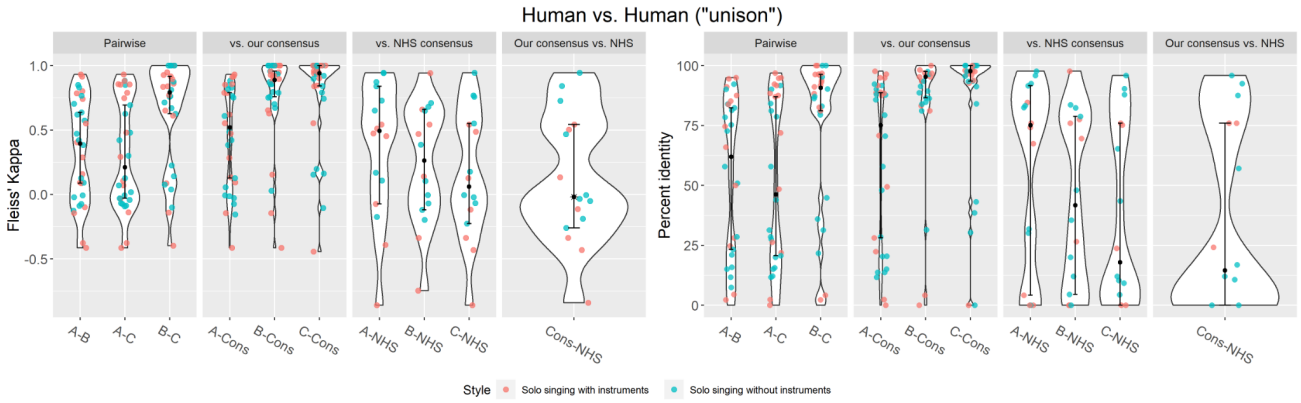

Figure S2. Agreement of human transcriptions not applying transposition.

J. SUMMARY OF DISAGREEMENT FACTORS OF LOW DISAGREEMENT SONGS

Song	Segmentation	Pitch	Both
NAIV-015	1	0	1
NAIV-048	0	0	1
NAIV-100	3	3	0
NAIV-117	0	4	0
T5431R27	0	0	3
T5482R03	0	3	0

Table S1. Qualitative classification of major disagreement factors of 19 pairs. The number indicates the count by segmentation disagreement, pitch disagreement or both factors.

K. AGREEMENT BETWEEN AUTOMATED METHODS AND INDIVIDUAL TRANSCRIBERS

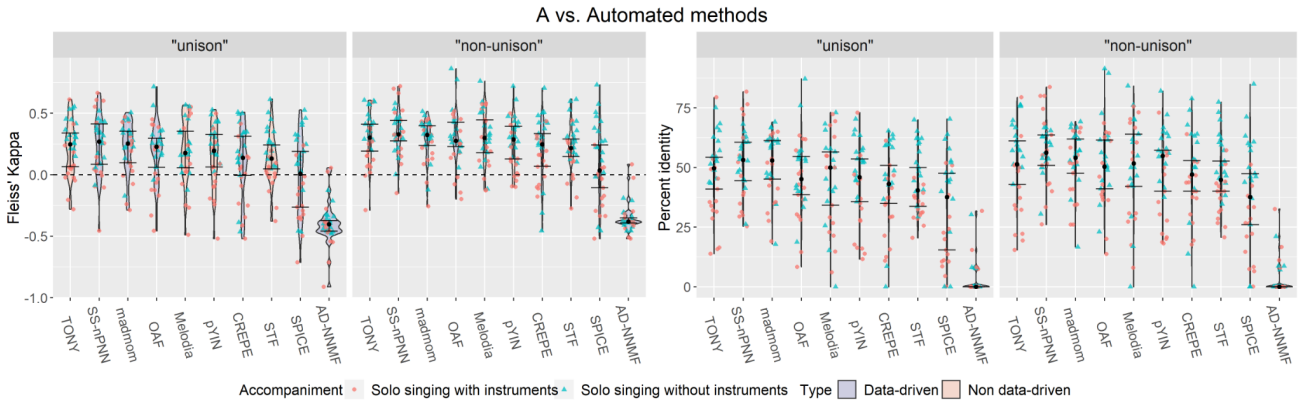

Figure S3. Pairwise agreement of automated methods vs. human ground-truth transcriptions.

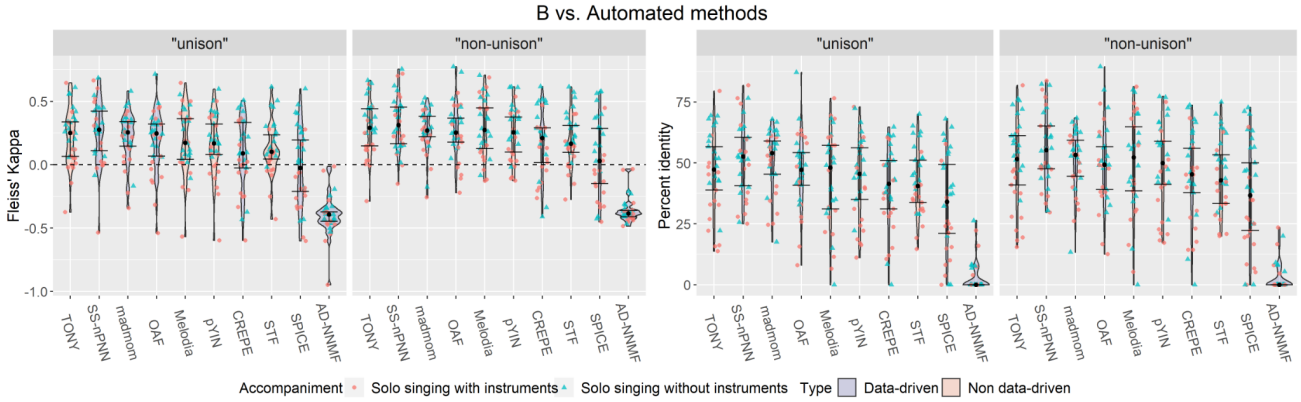

Figure S4. Pairwise agreement of automated methods vs. human ground-truth transcriptions.

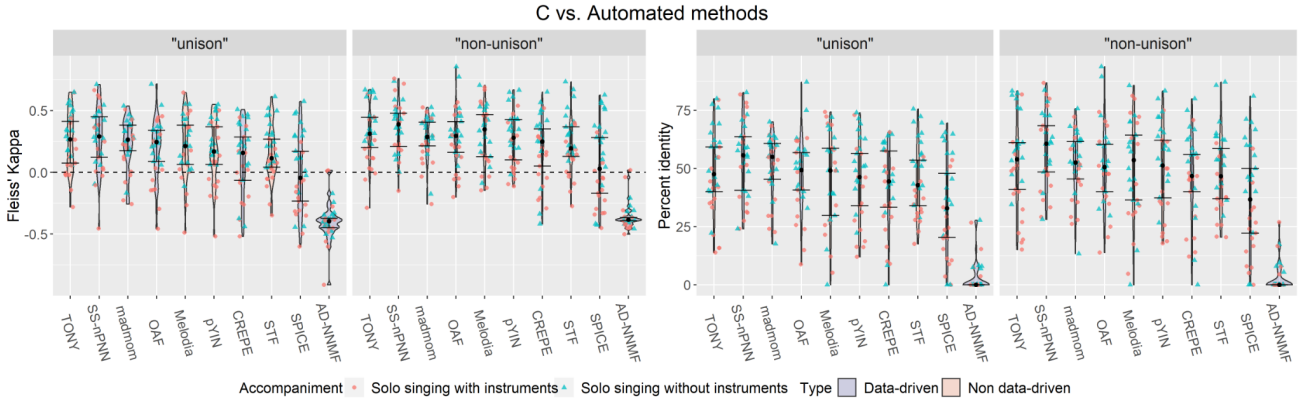

Figure S5. Pairwise agreement of automated methods vs. human ground-truth transcriptions.

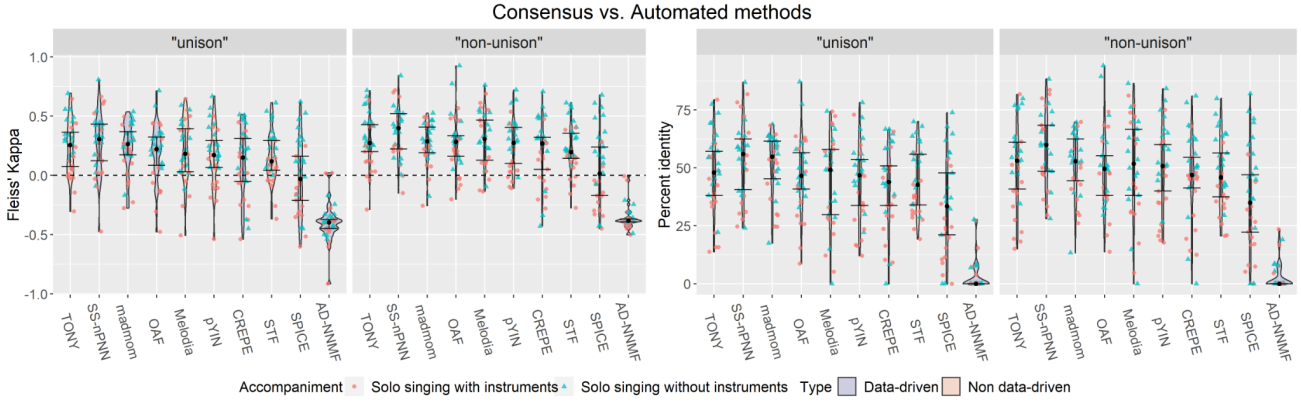

Figure S6. Pairwise agreement of automated methods vs. human ground-truth transcriptions.

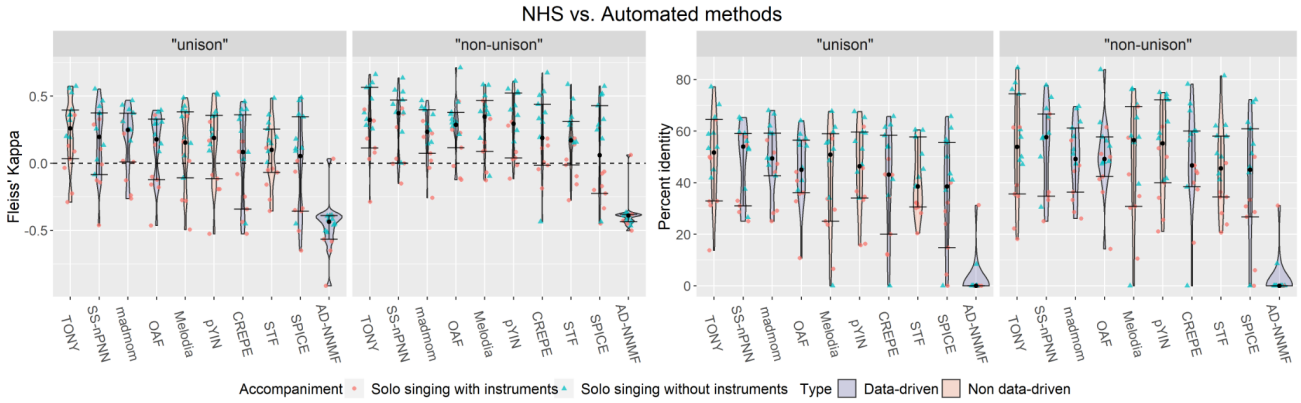

Figure S7. Pairwise agreement of automated methods vs. human ground-truth transcriptions.

L. RESULTS OF HYPOTHESIS TESTING

Median of κ = 0	p-value	α (BH)	ES (g)
Consensus vs. NHS	**<0.001**	0.010	0.5

Table S2. Result of the one-sample test. α (BH) is a threshold adjusted by the Benjamini–Hochberg step-up procedure

TONY vs.	p-value	α (BH)	$p(H_1\|X)$	ES (A)
AD-NMF	**<0.001**	0.005	**1.000**	0.985
CREPE	0.104	0.020	0.443	0.623
madmom	0.655	0.040	0.371	0.499
OAF	0.639	0.030	0.152	0.541
SPICE	**0.001**	0.015	**0.970**	0.732
SS-nPNN	0.923	0.045	0.145	0.462
Melodia	0.962	0.050	0.193	0.524
STF	0.210	0.025	0.214	0.613
pYIN	0.655	0.035	0.334	0.556

Table S3. Results of the two-sample tests. α (BH) is a threshold adjusted by the Benjamini–Hochberg step-up procedure.

M. NOTE LENGTHS OF NOTE SEQUENCES BY AUTOMATED METHODS

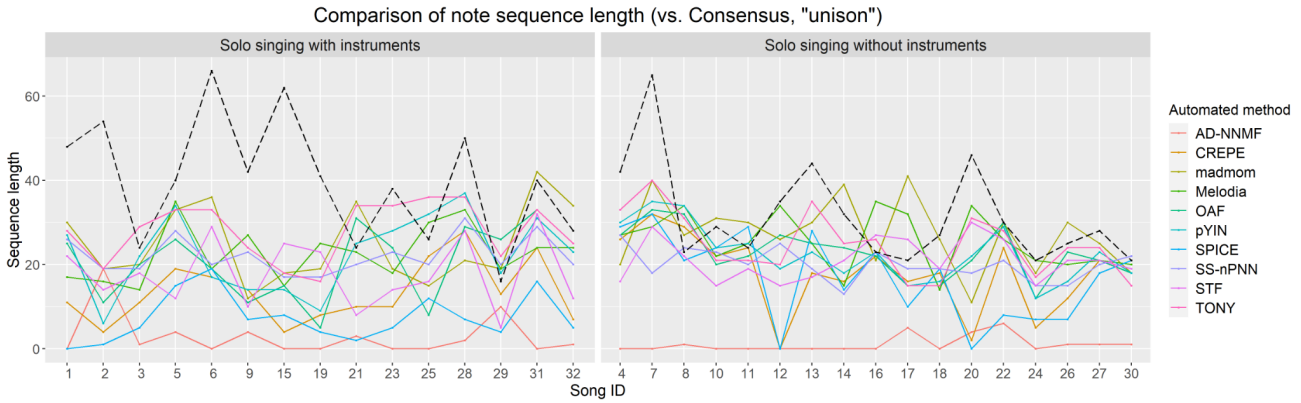

Figure S8. Lengths of note sequences by automated methods. The dashed line corresponds to the human note sequences, and the gap against that indicates that notes are segmented more or less than human transcription.

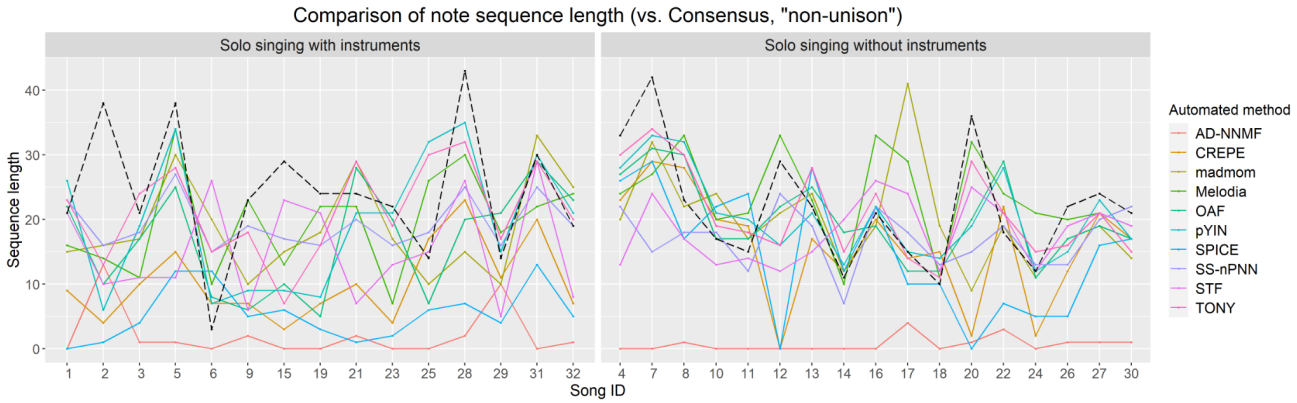

Figure S9. Lengths of note sequences by automated methods. The dashed line corresponds to the human note sequences, and the gap against that indicates that notes are segmented more or less than human transcription.

N. DIFFERENCE IN THE ORDER OF AGREEMENT SCORE BY SONG STYLE

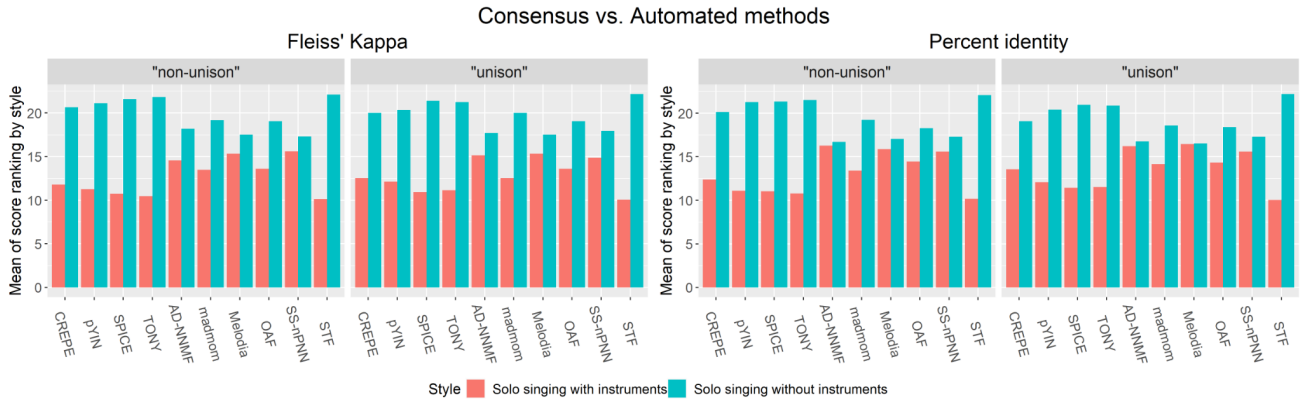

Figure S10. Difference of the average of ranking of scores by song styles. Scores of the 32 songs were ranked by descending order. The gap of average ranking indicates the automated method performed well for one style compared to the other.

O. FACTORS AND PATTERNS OF DISAGREEMENT BY LEVENSHTEIN DISTANCE

The below figures show varying patterns of disagreement among the note sequences of human and automated methods. We picked up 4 automated methods as representative samples. Furthermore, we chose the "non-unison" version to be able to evaluate the F0 prediction performance more directly.

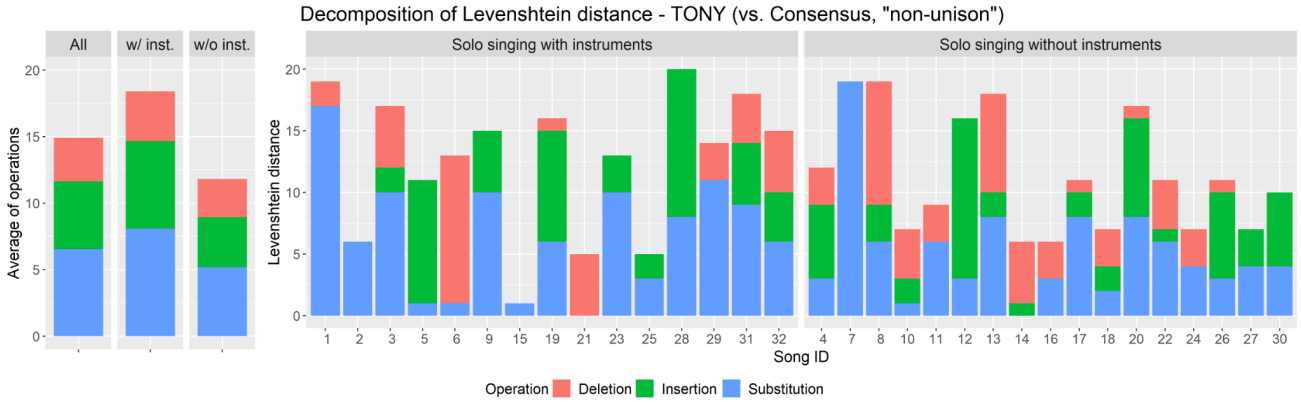

Figure S11. Type of disagreement decomposed by operation types.

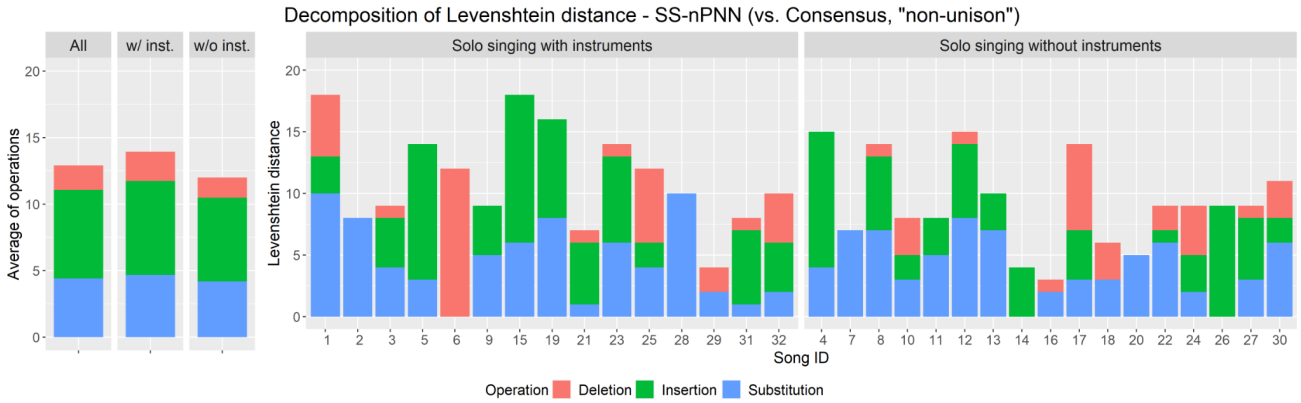

Figure S12. Type of disagreement decomposed by operation types.

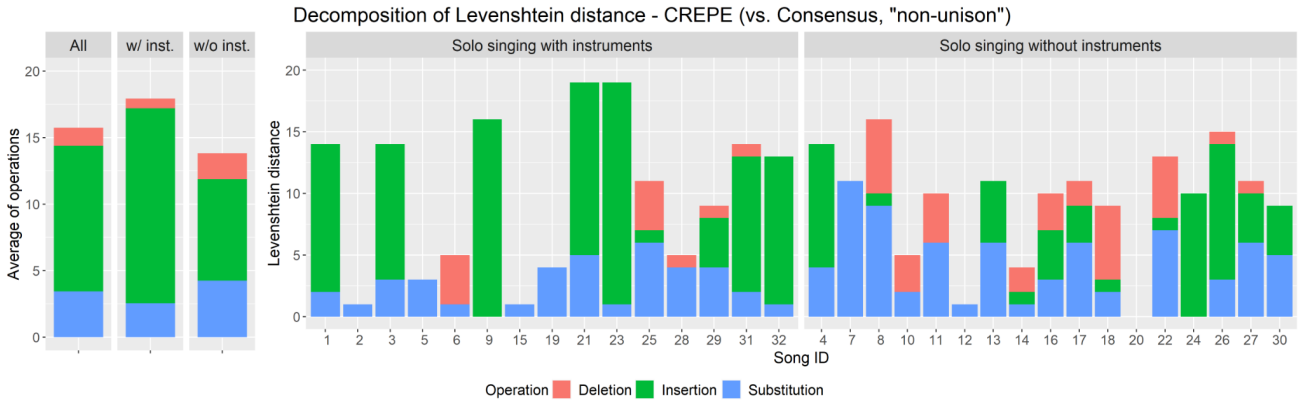

Figure S13. Type of disagreement decomposed by operation types.

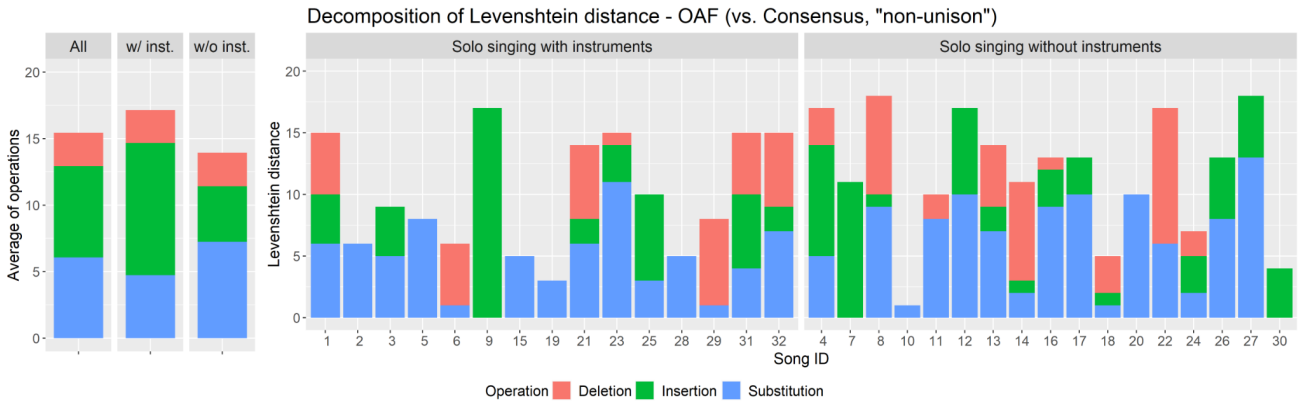

Figure S14. Type of disagreement decomposed by operation types.

P. SUMMARY OF TOP 10 OVERLAPPED BEST AGREEMENT RESULTS BY AUTOMATED METHODS

We first picked up the top 10 agreement songs in reference to our consensus note sequences from each automated method. After that, we further picked up the top 10 overlapping songs from that result.

Song	Song style	# of top-10 ranking in	Max κ	Automated method
NAIV-054	Solo singing without instruments	8	0.59	Melodia
NAIV-117	Solo singing without instruments	8	0.81	SS-nPNN
T5468R28	Solo singing without instruments	8	0.56	TONY
T5522R80	Solo singing without instruments	8	0.72	OAF
T5528R18	Solo singing with instruments	8	0.62	SS-nPNN
NAIV-021	Solo singing without instruments	7	0.56	TONY
NAIV-029	Solo singing with instruments	5	0.65	TONY

T5482R03	Solo singing with instruments	5	0.40	TONY
NAIV-075	Solo singing without instruments	4	0.47	madmom
T5421R17	Solo singing with instruments	4	0.67	SS-nPNN

Table S4. Results by the "unison" note sequence version.

Song	Song style	# of top-10 ranking in	Max κ	Automated method
NAIV-054	Solo singing without instruments	9	0.93	OAF
NAIV-104	Solo singing without instruments	8	0.58	CREPE
NAIV-117	Solo singing without instruments	8	0.84	SS-nPNN
T5468R28	Solo singing without instruments	8	0.67	TONY
T5522R80	Solo singing without instruments	7	0.77	OAF
T5528R18	Solo singing with instruments	7	0.70	SS-nPNN
NAIV-021	Solo singing without instruments	6	0.61	pYIN
NAIV-029	Solo singing with instruments	4	0.64	TONY
T5421R17	Solo singing with instruments	4	0.67	SS-nPNN
T5487R13	Solo singing with instruments	4	0.72	SS-nPNN

Table S5. Results by the "non-unison" note sequence version.

Q. **REFERENCES**

[58] I. Djurovic, and L. J. Stankovic, "An algorithm for the Wigner distribution based instantaneous frequency estimation in a high noise environment," *Signal Processing*, vol. 84, no. 3, pp. 631–643, 2004, doi: 10.1016/j.sigpro.2003.12.006.

[59] D. Gusfield, "Efficient methods for multiple sequence alignment with guaranteed error bounds," *Bulletin of Mathematical Biology*, vol. 55, no. 1, pp.141-154, 1993.

[60] G. Yona, *Introduction to Computational Proteomics*, Boca Raton, FL, USA: Chapman and Hall/CRC, 2010.

[61] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochemia Medica*, vol. 22, no. 3, pp. 276-282, 2012.

[62] W. R. Pearson, "An introduction to sequence similarity ("homology") searching," *Current Protocols in Bioinformatics*, vol. 42, no. 1, pp. 3.1.1-3.1.8, 2013, doi: 10.1002/0471250953.bi0301s42.

[63] A. N. Pettitt, "A two-sample Anderson-Darling rank statistic," *Biometrika*, vol. 63, no. 1, pp. 161-168, 1976.

[64] C. C. Holmes, F. Caron, J. E. Griffin, and D. A. Stephens, "Two-sample Bayesian nonparametric hypothesis testing," *Bayesian Analysis*, vol. 10, no. 2, pp. 297–320, 2015.

[65] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences, 2nd Ed.*, New York, NY, USA: Routledge, 1988.

[66] J. Ruscio, "A probability-based measure of effect size: Robustness to base rates and other factors," *Psychological Methods*, vol. 13, no. 1, pp. 19–30, 2008, doi: 10.1037/1082-989X.13.1.19.

[67] J. Cohen, "A power primer," *Psychological Bulletin*, vol. 112, no. 1, pp. 155-159, 1992.

[68] Y. Benjamini, and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 57, no. 1, pp. 289-300, 1995.