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

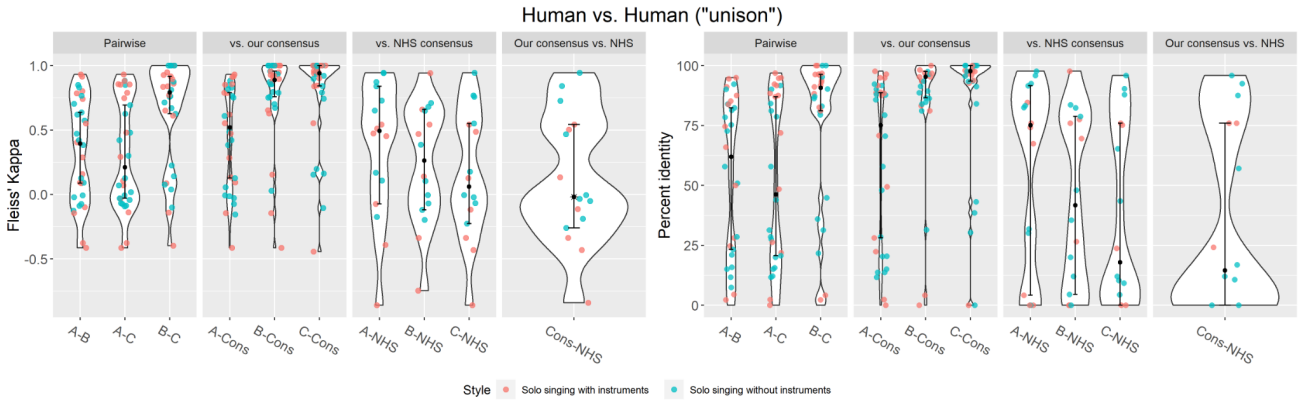

Figure S2. Agreement of human transcriptions not applying transposition.

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

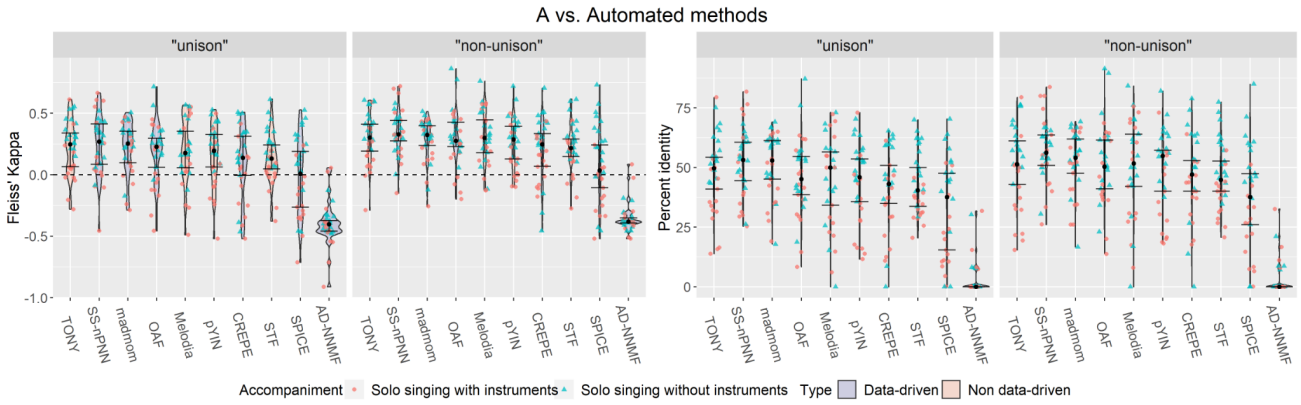

Figure S3. Pairwise agreement of automated methods vs. human ground-truth transcriptions.

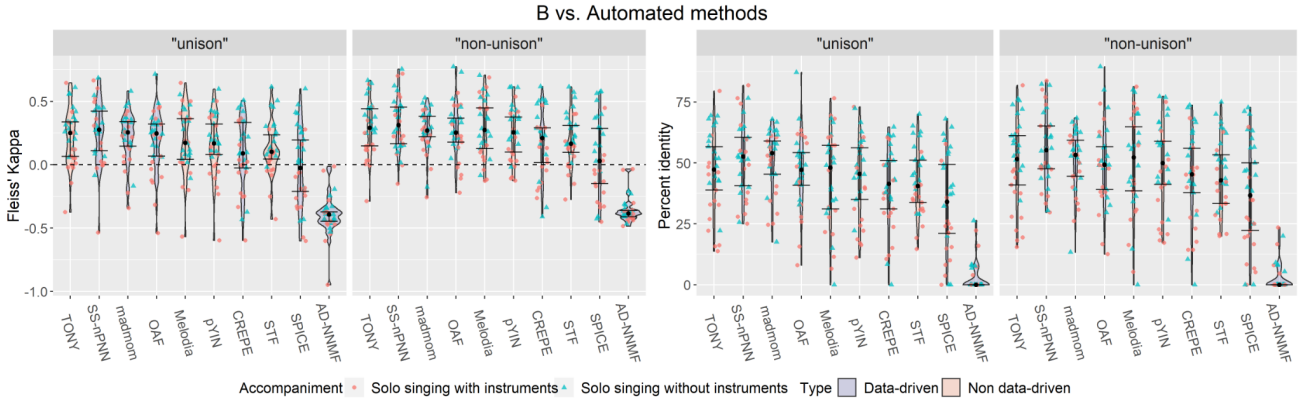

Figure S4. Pairwise agreement of automated methods vs. human ground-truth transcriptions.

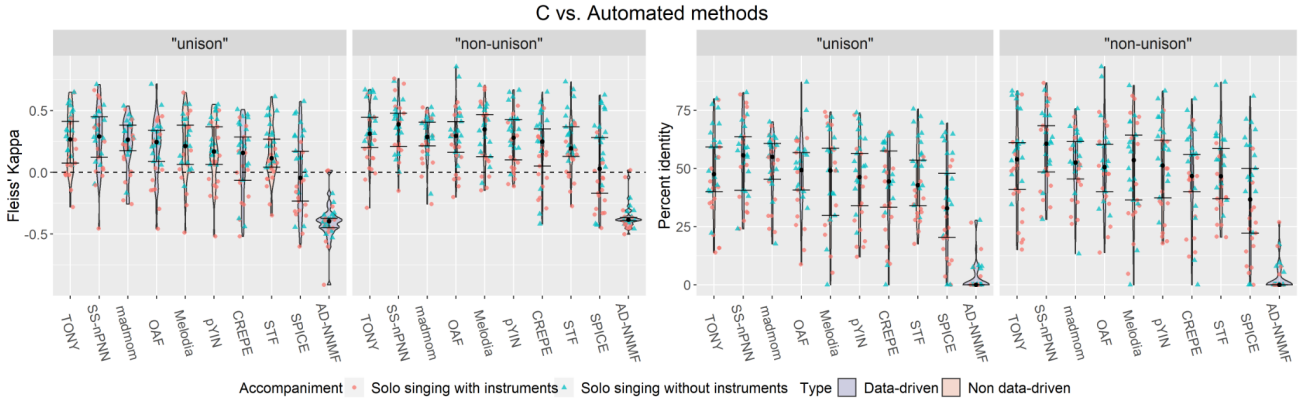

Figure S5. Pairwise agreement of automated methods vs. human ground-truth transcriptions.

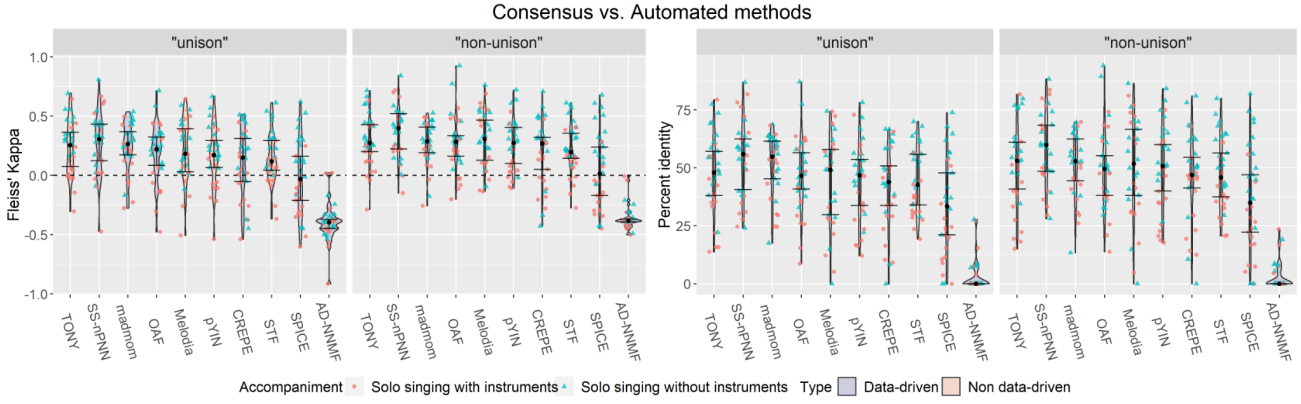

Figure S6. Pairwise agreement of automated methods vs. human ground-truth transcriptions.

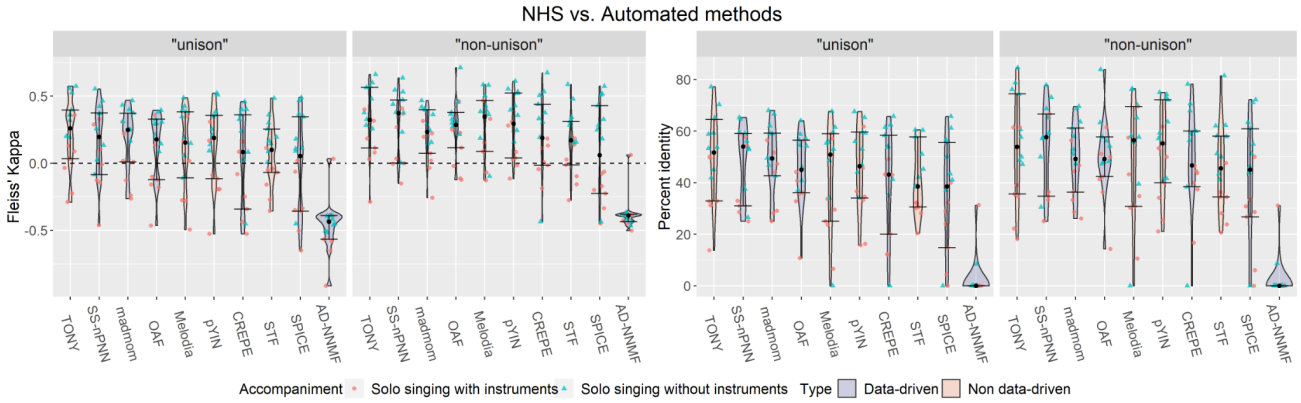

Figure S7. Pairwise agreement of automated methods vs. human ground-truth transcriptions.

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

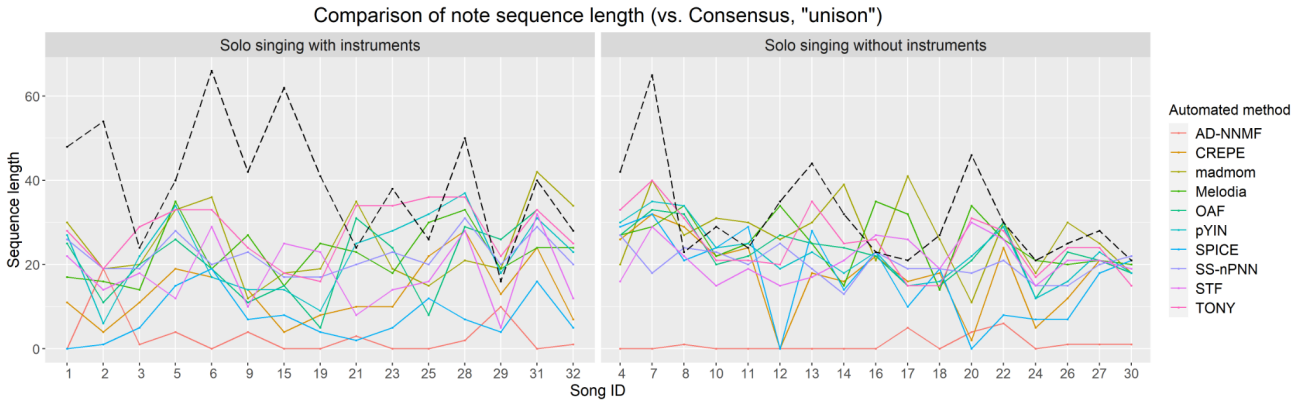

Figure S8. Lengths of note sequences by automated methods. The dashed line corresponds to the human note sequences, and the gap against that indicates that notes are segmented more or less than human transcription.

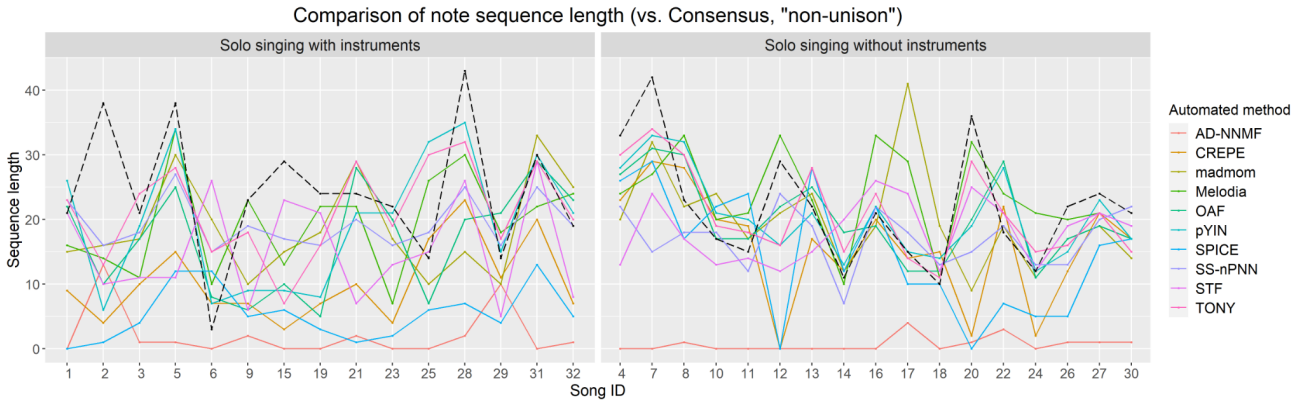

Figure S9. Lengths of note sequences by automated methods. The dashed line corresponds to the human note sequences, and the gap against that indicates that notes are segmented more or less than human transcription.

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

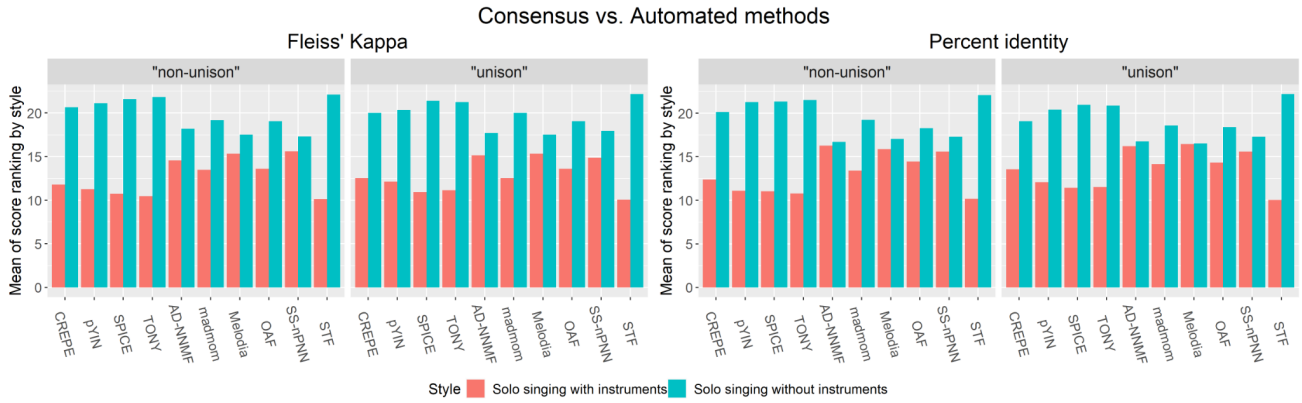

Figure S10. Difference of the average of ranking of scores by song styles. Scores of the 32 songs were ranked by descending order. The gap of average ranking indicates the automated method performed well for one style compared to the other.

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.

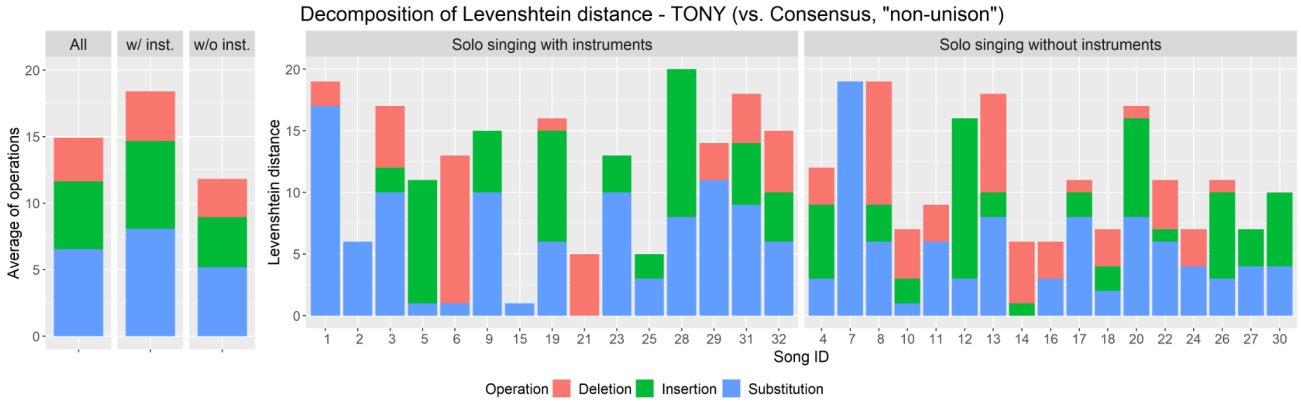

Figure S11. Type of disagreement decomposed by operation types.

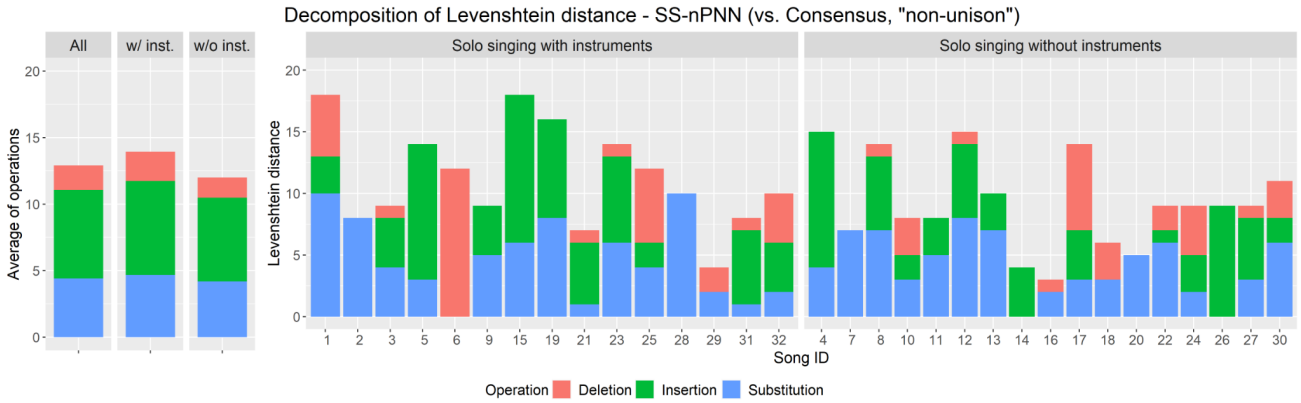

Figure S12. Type of disagreement decomposed by operation types.

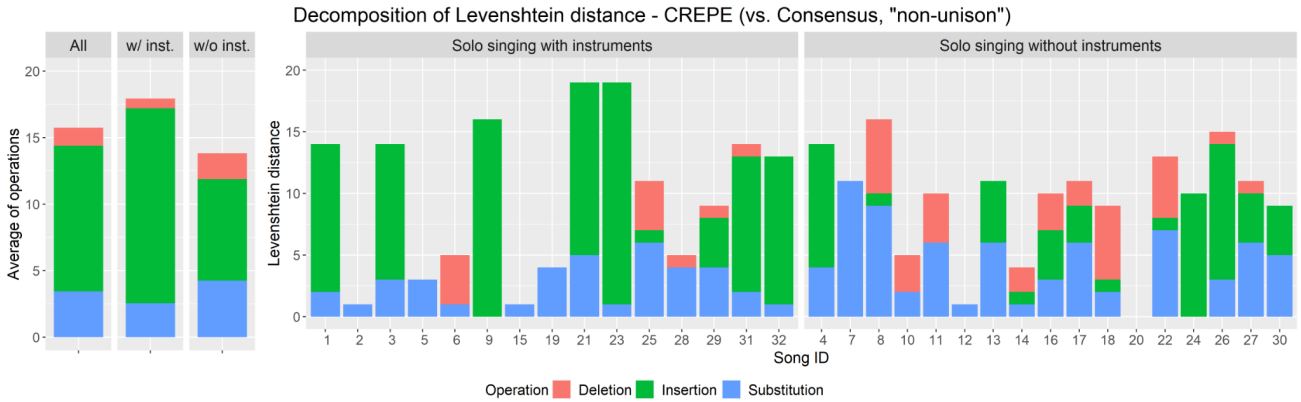

Figure S13. Type of disagreement decomposed by operation types.

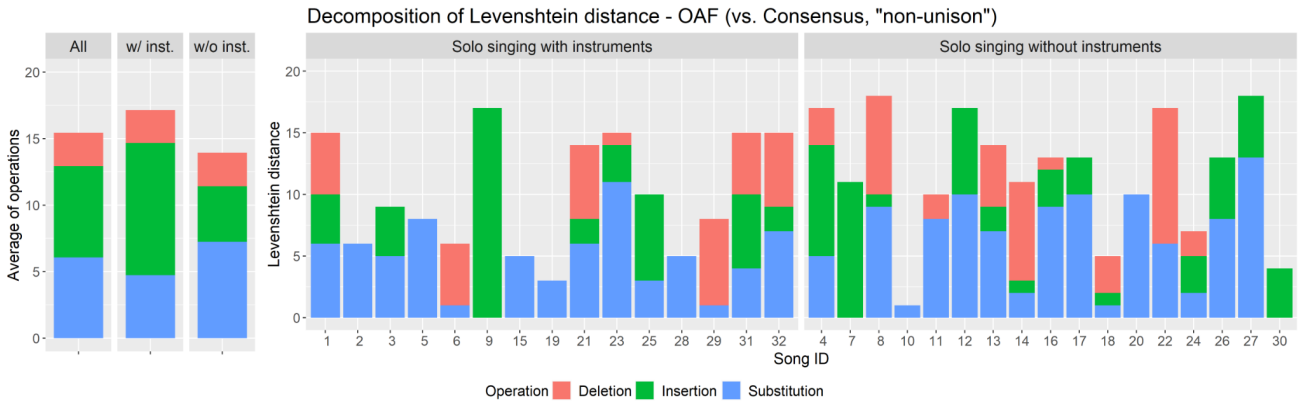

Figure S14. Type of disagreement decomposed by operation types.

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.