Cut Circle: Intonation and Timing

# Method

## Procedure

The choir sung four Renaissance pieces (Table 1) in three different conditions three times on three different days, except for Guillaume Dufay’s “Missa Ecce ancilla domini”, which was only sung twice due to a recording error on the first day. Each condition and work comprises one take, except for the two pieces by Josquin where always combined. This resulted in 24 takes in total (see Table A1 and A2 in the appendix for a full overview).

The three experimental conditions were related to the physical setup of singers: “far apart (no touch)” (*far-apart*), “close (no touch)” (*no-touch*), and “close (with touching)” (*touch*). The scores were projected via a projector. See Lange et. al. (2022) for more information on the setup details.

The two works by Josquin consisted of a single parts, whereas the Dufay masses consists of three parts each, which were partly repeated in a take in order to achieve a recording duration of at least 6 minutes for the physiological measurements. The same holds true for Josquin’s “D’ung Altre”.

All works are set in four-part harmony (named here soprano, alto, tenor, and bass) except Dufay’s Agnus II which features only soprano and alto. All voices were sung by two singers singing in unison except for a few final notes, which were split in two (*dufay-agnus1*, *dufay-kyrie1*) or, on a single occasion, in three notes (*dufay-kyrie1*).

All singers were equipped with headsets and recorded separately on each take (soprano: headsets 7 and 8, alto: headsets 1 and 2, tenor: headsets 5 and 6, bass: headsets 3 and 4). To ease the subsequent transcription step and because they represent closed musical units, each repetition of each part of each work was extracted from the recording. In total, there were 93 of these sections with 8 or 4 audio tracks by headset derived from them, for a total of 672 audio tracks.

The audio tracks were transcribed into note tracks with the help of **Tony** software (REF TONY) and manually corrected and optimized when the algorithm gave erroneous or sub-optimal results.

Subsequently, the note tracks with exact onsets, duration, and pitches (measured in Hz) were imported into **Sonic Visualiser** and in manually annotated with note labels, which were generated from digital scores of all works with the help of **music21**. During this step, extra and misplaced notes were also annotated by the transcribing musicologist (A.S.). All annotated note tracks were exported in CSV format and imported into R for further analysis, along with various metadata. This resulted in 672 note tracks with a total of 65280 tone events. For further processing, a variant of the data set was created by filtering notes that were annotated as errors, resulting in a remaining set of 64215 note events. A complete overview of note tracks, sections, parts and number of note events can be found in the appendix (Table A2).

Table 1. Works used in the study

| ID | Work |
| --- | --- |
| dufay-agnus1 | Guillaume Dufay. Missa Ecce ancilla domini, Agnus I (Duf1003e) |
| dufay-agnus2 | Guillaume Dufay. Missa Ecce ancilla domini, Agnus II (Duf1003e) |
| dufay-agnus3 | Guillaume Dufay. Missa Ecce ancilla domini, Agnus III (Duf1003e) |
| josquin-virgo | Josquin des Prez, Virgo Prudentissima (NJE 25.12) |
| josquin-dung | Josquin des Prez, Missa D'ung Aultre Amer (NJE 7.3) |
| dufay-kyrie1 | Guillaume Dufay. Missa Ecce ancilla domini, Kyrie I (Duf1003a) |
| dufay-kyrie2 | Guillaume Dufay. Missa Ecce ancilla domini, Kyrie II (Duf1003a) |
| dufay-gloria | Guillaume Dufay, Missa Ecce ancilla domini, Gloria (Duf1003a) |

A digital representation of the scores consisting of nominal MIDI pitch and nominal onset (in units of smallest duration, which was the quarter note length) were also imported and joined to the note track data. From this final representations several performance accuracy measures were derived pertaining to intonation and timing.

## Performenc Indicators

### Intonation

#### Preprocessing to account for pitch drift and global tuning

It turned out that no take was sung according to perfect 440 Hz standard tuning. Furthermore, it is well-known that choirs can exhibit pitch drift, which would distort the measurement of deviation from nominal pitch. In order to account for both effects, we removed any possible linear trend from the overall series of pitch deviations across all voices, as tuning and drift can be considered a global phenomenon on the choir level. Sections (parts and repetitions) were treated separately.

The raw pitch values in Hz of the note tracks were converted to fractional MIDI pitch based on 440 Hz concert tuning (a’ = 440 Hz = MIDI pitch 69). The deviations of pitch values from nominal pitch include thus a global deviation from standard tuning as well as possible linear trends. A simple linear regression model over onsets and pitch deviation was used to removed drift and tuning bias from the pitch deviations in one step, via substituting raw pitch deviations with the residual values from the linear regression. We used furthermore the pitch drift slopes as an extra indicator for performance accuracy.

#### Pitch accuracy and precision

We derived a set of performance indicators from the pitch deviations. First, we defined the mean absolute pitch error (MAPE) as the singer-wise mean of absolute pitch deviations per single note track (section). Additionally, we used the standard deviation of singer- and section-wise pitch deviation as a measure for mean pitch precision (MPP). Since both measures are always positive and not normally distributed, we used the negative logarithms of both measures for modeling purposes to ensure near-gaussian error term. We well call these derived measures mean pitch accuracy and mean pitch precision in the following, for which now higher values indicate more precise and more accurate intonation.

One important point has to be discussed here. In using the deviation from nominal MIDI pitch values taken from the scores, we implicitly assume that the singers were using 12-tone equal temperament for their intonation, which is, strictly speaking, not necessarily a valid assumption, even though previous studies (Mauch et al, Fischinger et al.) showed that the actually employed intonation strategy cannot be reliably inferred from real singing data (though a small preference for 12-TET was observed in said studies.) One reason is that the theoretical differences between 12-TET and, say, just intonation are in same order of magnitude as the intonation accuracy of professional singers (and much smaller than the accuracy of amateur singers). This is even further complicated by global pitch drifts, which is also in the same order of magnitude. Since we want mainly to compare different condition here, the assumption of a 12-TET intonation seems not particularly restrictive, if we can assume that the choir uses basically the same intonation strategy across all pieces and conditions, which seems to be a safe assumption. The potential systematic error stemming from the assumption of 12-TET intonation when in fact another strategy is used could result then, at least theoretically, in constant bias. The total mean absolute pitch error is 13.6 cents across all singers, which is smaller than the Pythagorean comma of about 23.46 cents, indicating that the assumption of 12-TET intonation strategy seems justified.

That being said, we were also defining a third measure of intonation quality which is independent from any assumptions of intonation strategy as it compares the pitch deviation within a voice as we have two singers per voice singing the same pitches. These measure is called inner-voice MAPE defined as the section-wise mean of absolute pitch differences between the two singers of a voice. Likewise, we define also a inner voice MPP, as the standard deviation of pitch differences between singers of the same voice within a section. For the same reasons as mentioned above, we actually used the negative logarithm for numerical analyses.

Finally, we also used the slope of the overall pitch drift as global indicator for performance accuracy (even though it can be argued that pitch drift is not a strong indicator for this but mainly a result of intonation adaptations.)

### Timing

The second class of performance indicators deals with note timing. In contrast to intonation, we did not compare here performed note timing to the nominal timing as prescribed by the score, as perfectly strict timing is not a common stylistic goal in Renaissance music performances. Instead, we only used within-voice and across-voice synchronization measures. We define mean onset precision (MOP) as the section-wise mean of standard deviations for the onsets on synchronization points, which are points in the score where at least two voices have a common note onset (this amounts to 76.2% of all note events). We filtered outliers (based on the usual 1.5 \* IQR criterion) as gross singing errors, as they have a large influence on these indicators, which is not desired.

Inner voice onset precision is similarly defined as the section-wise mean standard deviation of onset differences between the singers within a voice, where we filtered events that were more than 300 ms apart, as these can be considered singing errors, since the maximum tempo of all sections was 250 ms for quarter notes, so a deviation of more than a quarter note should not be considered as micro-timing anymore.

Additionally, we define mean onset error (MOE) as the section-wise average of absolute onset differences. Again, for linear regression models we used the negative logarithm of these values and use the term mean onset precision.

We also analyzed overall tempo and tempo drift as further performance indicators for timing. Tempo was measured as the ratio of performed inter-onset intervals (IOI) to nominal inter-onset intervals, where we only used IOIs with nominal values of 2 or 4 metrical units (tatums), which amount to half and full notes. Global tempo drift was estimated using the slope of a linear regression model of measured quarter duration against onset.

### Singing errors

Last but not least, the ostentatious singing errors can be used as performance indicators. Here we used the amount of annotated erroneous tones, which included wrong pitches and pitches that were interrupted. Additionally, we also counted pitch deviations larger than a semitone as a singing errors, even when consonant in the actual context and onset differences larger than a quarter note length.

# Results

## Pitch accuracy and precision

We used linear mixed regression models to estimate the effect of condition (*touch*, *no-touch*, *far-apart*) on mean pitch accuracy and precision, where we used day of performance, piece, and singer as random effects. As can be seen in Figs. 1a and 1b, and corroborated by a permutation test (coin::indepence\_test), there are significant differences for all four variables. Two models for mean pitch accuracy and mean pitch precision were calculated (lmerTest package for R), which can be found in Tab. 2. There were significant difference for all indicators and variables except for the *no-touch* condition for mean pitch accuracy (reference category is *far-apart*). The *touch* category showed lower pitch accuracy and pitch precision than the *no-touch* condition, which in turn showed lower than accuracy than the *far-apart* condition. Nakawa’s conditional and marginal R^2 were .441 and .008 for mean pitch accuracy and .402 and .022 for pitch precision. Hence, effects for pitch precision are generally stronger than for pitch accuracy. Furthermore, mean pitch accuracy and precision are strongly correlated (r(334) = .911), indicating that there is no systematic bias in pitch heights: all pitch errors are due to random variations around the true (nominal) pitch.

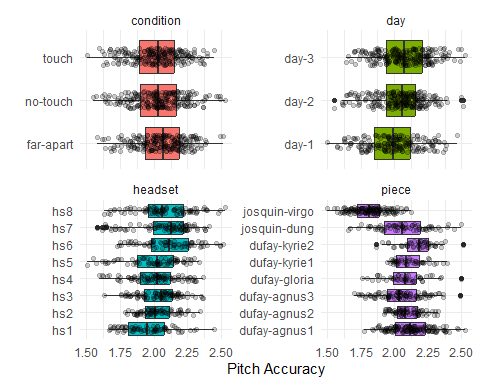


Figure 1a. Pitch accuracy for condition, day, piece and headset (singer)

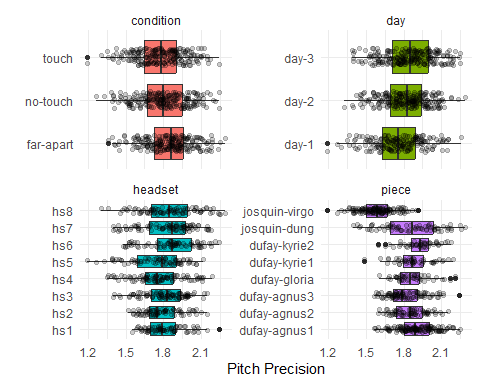


Figure 1b. Pitch precision for condition, day, piece and headset (singer)

Table 2. Mixed linear regression models for pitch accuracy and pitch precision.

| Indicator | Term | beta | p |
| --- | --- | --- | --- |
| Pitch Accuracy | (Intercept) | 2.077 | 0.000\*\*\* |
|  | no-touch | -0.021 | 0.097. |
|  | touch | -0.039 | 0.002\*\* |
| Pitch Precision | (Intercept) | 1.865 | 0.000\*\*\* |
|  | no-touch | -0.046 | 0.001\*\*\* |
|  | touch | -0.066 | 0.000\*\*\* |

The linear regression model for within pitch accuracy and precision can be found in Tab. 3. Here, we see only significant effects of inner pitch accuracy but not for inner pitch precision, indicating that there is systematic bias between two voices.

Table 3. Linear Regression models for within pitch accuracy and pitch precision.

| Indicator | Term | beta | p |
| --- | --- | --- | --- |
| Pitch Accuracy | (Intercept) | 1.884 | 0.000\*\*\* |
|  | no-touch | -0.090 | 0.000\*\*\* |
|  | touch | -0.105 | 0.000\*\*\* |
| Pitch Precision | (Intercept) | 1.528 | 0.000\*\*\* |
|  | no-touch | -0.072 | 0.101 |
|  | touch | -0.067 | 0.128 |

In order to get estimates for the relative effect sizes of intonation quality, we resorted to standardization of global and inner pitch accuracy and pitch precision by day, headset, and piece followed by calculation Tukey contrasts, which give then approximate effect sizes for the single contrast. Effects are generally of low to medium size, the largest can be found for *touch* vs. *far-part* conditions for inner pitch accuracy with d = -.819. Generally, the largest effect sizes can be found for this pair of conditions for all four performance indicators. However, the absolute effect sizes between conditions in terms of absolute pitch errors are less than 1 cent (average 0.4 cents, range -2.1 to 3.7 cents for *touch* vs. *far-apart*), which is way less than any pitch discrimination threshold.

Table 4. Approximate effect sizes global and inner pitch accuracy and pitch precision.

| Indicator | Contrast | Cohen's d | p |
| --- | --- | --- | --- |
| Pitch Accuracy | No Touch --. Far Apart | -0.124 | 0.290 |
|  | Touch --. Far Apart | -0.281 | 0.002\*\* |
|  | Touch --. No touch | -0.157 | 0.137 |
| Pitch Precision | No Touch --. Far Apart | -0.358 | 0.000\*\*\* |
|  | Touch --. Far Apart | -0.500 | 0.000\*\*\* |
|  | Touch --. No touch | -0.142 | 0.184 |
| Within Pitch Accuracy | No Touch --. Far Apart | -0.670 | 0.000\*\*\* |
|  | Touch --. Far Apart | -0.824 | 0.000\*\*\* |
|  | Touch --. No touch | -0.154 | 0.322 |
| Within Pitch Precision | No Touch --. Far Apart | -0.519 | 0.000\*\*\* |
|  | Touch --. Far Apart | -0.583 | 0.000\*\*\* |
|  | Touch --. No touch | -0.065 | 0.833 |

### Pitch errors

For gross pitch errors, as defined above, We calculated a mixed logistic regression with day, piece, and headset as random effects and condition as fixed effects. The results can be found in Table 5. Both the *touch* and the *no-touch* condition showed a significant increase of error rates compared to the *far-apart* condition with a 55% and 59% increase of odd ratios, resp.

Table 5. Mixed logistic regression models for singing errors.

| Term | beta | exp(beta) | p |
| --- | --- | --- | --- |
| (Intercept) | -5.688 | 0.003 | 0.000\*\*\* |
| no-touch | 0.441 | 1.554 | 0.001\*\* |
| touch | 0.467 | 1.595 | 0.001\*\*\* |

### Pitch drift

We used the slope of overall pitch drift per section as another performance indicator and subjected it to a mixed linear regression with condition as fixed effect and day and piece as random effect. The model did not show significant results for condition, though there was considerable variation in pitch drift with respect to sections, mostly showing downward drift (78.8 % of all cases). About of 24.6% the slopes were significantly different from 0 on the 5% level, indicating that pitch drift is a very common phenomenon.

Total drift across sections is, however, small with -9.7 cents on average and a range from -66.1 to 22.9 cents.

## Timing

### Onset precision

The distribution of onset precision between voices for condition, day, and piece can be seen in Fig. 2a, and those for within voices in Fig. 2b. To check for main effects of condition, piece, day, and voice type (for inner onset precision) we again used a permutation test (coin::indepence\_test), which only showed effects for piece for mean onset precision and for piece and voice type for inner onset precision. However, the mixed linear regression models for both types of onset precision, using condition as fixed effect and day and piece (day, piece, and voice type, resp.), did show one significant effect for the *touch* condition in the case of mean onset precision across voices. This is has a medium relative effect size of d = -.662, i.e., lower onset precision for *touch* compared to *far-apart* (again calculated using standardization across random effects and then using Tukey contrasts). However, the absolute effect size is only d = -0.0035, i.e., about 4 ms, way below any onset discrimination threshold.

Table 6. Mixed linear regression models for mean onset precision between voices.

| Indicator | Term | beta | p |
| --- | --- | --- | --- |
| Mean Onset Precision | (Intercept) | 3.072 | 0.000\*\*\* |
|  | no-touch | 0.008 | 0.463 |
|  | touch | 0.043 | 0.000\*\*\* |

Table 7. Mixed linear regression models for mean onset precision within voices.

| Indicator | Term | beta | p |
| --- | --- | --- | --- |
| Mean Onset Precision | (Intercept) | 2.805 | 0.000\*\*\* |
|  | no-touch | -0.021 | 0.245 |
|  | touch | -0.009 | 0.614 |

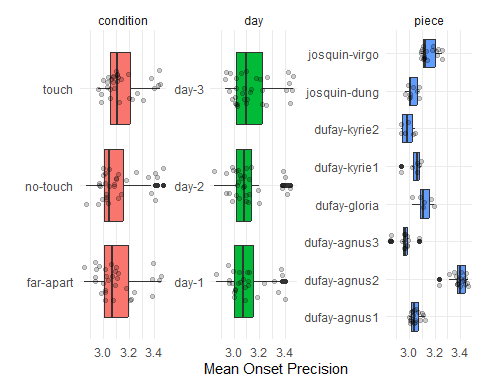


Figure 2a. Mean onset precision for condition, day, and piece

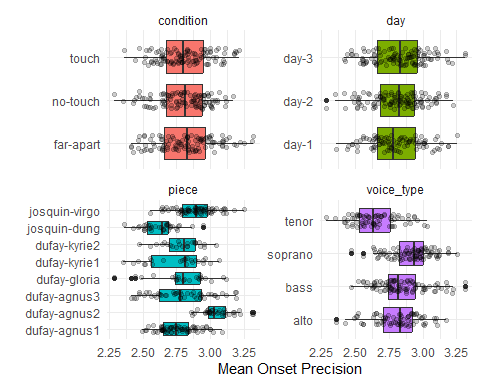


Figure 2b. Inner onset precision for condition, day, piece and voice type

Table 8. Approximate effect sizes global and inner onset precision.

| Indicator | Contrast | Cohen's d | p |
| --- | --- | --- | --- |
| Onset Precision (Across) | No Touch --. Far Apart | 0.237 | 0.488 |
|  | Touch --. Far Apart | 0.870 | 0.000\*\*\* |
|  | Touch --. No touch | 0.633 | 0.008\*\* |
| Onset Precision (Within) | No Touch --. Far Apart | -0.150 | 0.406 |
|  | Touch --. Far Apart | -0.080 | 0.773 |
|  | Touch --. No touch | 0.070 | 0.821 |

### Timing Errors

We analyzed timing errors, again across and within voices. Across voices, we counted any synchronization point as an timing error if the onset precision (standard deviation of onsets) at these points is an outlier in the distribution of all onset precisions (using the usual 1.5 \* IQR criterion). Subsequently, we calculated a mixed logistic regression model for errors with condition as fixed effect and day and piece as random effects. The results can be found in Table 8. No significant effect could be found, though, the p values for the *touch* condition is very close to the 5% significance level. Interestingly, the coefficient is negative, indicating **fewer** errors in the *touch* condition.

Table 8. Mixed logistic regression models for timing error between voices.

| Term | beta | exp(beta) | p |
| --- | --- | --- | --- |
| (Intercept) | -3.211 | 0.040 | 0.000\*\*\* |
| no-touch | 0.121 | 1.128 | 0.309 |
| touch | -0.249 | 0.780 | 0.053. |

Within voice, timing errors were defined as onset differences greater than a fixed threshold of 300 ms (maximal quarter note duration) and used a logistic regression for so defined errors with condition as fixed effect and day, piece, and voice type as random effects. Results can be found in Table 9. Here, the *no-touch* condition became significant with a 116% increase in odd ratios

Table 9. Mixed logistic regression models for timing error within voices.

| Term | beta | exp(beta) | p |
| --- | --- | --- | --- |
| (Intercept) | -5.831 | 0.003 | 0.000\*\*\* |
| no-touch | 0.771 | 2.163 | 0.000\*\*\* |
| touch | -0.067 | 0.935 | 0.767 |

### Tempo and tempo drift

Lastly, we also used the linear tempo trend as well as absolute tempo as performance indicators. To this end, We again used the negative logarithm of linear slopes, negative logarithm of total tempo change and mean tempo and calculated mixed linear models with condition as fixed effect and day and piece as random effects. Results can be found in Table 10. Except for significant non-zero intercepts, indicating that tempo drift was a common phenomenon, only the mean tempo of the *touch* condition compared to the *far-apart* condition became significant showing a negative sign, which indicates generally slower tempo in the *touch* condition. The relative effect size is medium with d = .688, but the absolute effect is at most 3.8 % or 5 ms, again, way below any perceptual threshold for onset perception.

Table 10. Mixed linear regression models for tempo slope, tempo change, and mean tempo

| Indicator | Term | beta | p |
| --- | --- | --- | --- |
| Neg. logarithm of absolute tempo slope | (Intercept) | 8.345 | 0.000\*\*\* |
|  | no-touch | 0.080 | 0.434 |
|  | touch | 0.128 | 0.214 |
| Mean quarter length duration | (Intercept) | 0.202 | 0.000\*\*\* |
|  | no-touch | 0.000 | 0.920 |
|  | touch | -0.003 | 0.001\*\*\* |
| Neg. logarithm of absolute tempo change | (Intercept) | 3.823 | 0.000\*\*\* |
|  | no-touch | 0.070 | 0.499 |
|  | touch | 0.132 | 0.205 |

# Discussion

The results draw a rather clear picture as thhe experimental conditions had some influence on the performance, mostly in the pitch domain. Generally, the *touch* condition performed worse than the *no-touch* which in turn performed worse than the *far-apart* condition, the latter being the normal (concert) setup for the choir. The interpretation is, contrary to our original hypothesis, that being closer or even touching each other has a slight deteriorating effect on performance, probably, because the singer are not accustomed to it. This untypical setup likely leads to distraction and possibly intermittent loss of focus. The *touch* and *no-touch* condition both breach the normal personal comfort zone while the *touch* condition even introduces physical contact, which likely does not happen often in this professionally-related, mixed-gender choir. That the performances got better over time (i.e., by days of recording) is further evidence that the observed effect might be mostly a matter of breaking habits. The two closer, unusual setups particularly provoked singing errors (general mistakes and in the pitch domain but not in regard to timing), and sometimes the choir members can even be heard laughing on tape. This further indicates that the slightly worse performece is due to getting distracting by the performance situation. Finally, many difference between the *touch* and *no-touch* condition are not significant which provides further evidence that this is an effect of either the unfamiliarity of the closer setups or the breach of individual comfort space.

However, one would expect then also a higher tempo in the *touch* condition, due to higher levels of arousal, but the opposite was the case, though with a very tiny effect. We can only speculate here on the reason for this. For instance, the slowing down might be on account of the conductor counteracting the perceived higher arousal.

Compared to monastic choirs in the medieval times, where touching during singing might be very common, if not the standard setup, or at least usual for singing lessons, the closer condition in our experiment means deviation from the norm. In this regard, the performance of this modern professional choir can hardly be fairly compared with medieval monks.

However, it has to be noted that the effect size are basically minuscule, particularly, when measured in absolute terms. The relative effects are low to medium but this is mainly an indicator of the overall high precision of the choir. Generally, all performance differences might be well below perceptual thresholds and, hence, probably unnoticeable by the audience. However, one might speculate that the close and *touch* conditions could have more detrimental effects on a less professional choir.

There is one observation that runs contrary to the drawn-out picture. The *no-touch* condition had an elevated number of timing errors compared to the *far-apart* setup instead of *touch* condition as could expected from the general pattern. Here, touching each other might, in fact, have helped keeping errors at bay despite the assumed distraction. But this is mere conjecture. The overall effect is very small and might be just spurious. Nonetheless, it should be noted that timing, in general, was much less affected by the experimental conditions in contrast to intonation, which might be in agreement with the observed higher levels of breathing synchronization in the *touch* condition (Lange et. al, 2022).

# Appendix

Table A!. Overview of analytical units.

| Element | Number |
| --- | --- |
| Works | 4 |
| Parts/Pieces | 8 |
| Takes | 24 |
| Sections/Audio Tracks | 93 |
| Singers/Headsets | 8 |
| Note Tracks | 672 |
| Voice Types | 4 |
| Conditions | 3 |
| Recording days | 3 |
| All tone events (all) | 65,280 |
| Tone events (excluding pitch errors) | 64,215 |

Table A2. Basic statistics of the single takes.

| Take | Condition | Piece | Repetition | No. Notes | No. Voices |
| --- | --- | --- | --- | --- | --- |
| take05 | touch | dufay-agnus1 | 1 | 609 | 4 |
|  |  | dufay-agnus2 | 1 | 429 | 2 |
|  |  | dufay-agnus3 | 1 | 601 | 4 |
|  |  | dufay-agnus1 | 2 | 617 | 4 |
|  |  | dufay-agnus2 | 2 | 429 | 2 |
|  |  | dufay-agnus3 | 2 | 606 | 4 |
| take06 | no-touch | dufay-agnus1 | 1 | 624 | 4 |
|  |  | dufay-agnus2 | 1 | 430 | 2 |
|  |  | dufay-agnus3 | 1 | 608 | 4 |
|  |  | dufay-agnus1 | 2 | 612 | 4 |
|  |  | dufay-agnus2 | 2 | 427 | 2 |
|  |  | dufay-agnus3 | 2 | 606 | 4 |
| take07 | far-apart | dufay-agnus1 | 1 | 618 | 4 |
|  |  | dufay-agnus2 | 1 | 434 | 2 |
|  |  | dufay-agnus3 | 1 | 613 | 4 |
|  |  | dufay-agnus1 | 2 | 610 | 4 |
|  |  | dufay-agnus2 | 2 | 427 | 2 |
|  |  | dufay-agnus3 | 2 | 614 | 4 |
| take08 | touch | josquin-dung | 1 | 479 | 4 |
|  |  | josquin-virgo | 1 | 1,003 | 4 |
|  |  | josquin-virgo | 2 | 889 | 4 |
| take09 | no-touch | josquin-dung | 1 | 491 | 4 |
|  |  | josquin-virgo | 1 | 1,033 | 4 |
|  |  | josquin-virgo | 2 | 1,001 | 4 |
| take10 | far-apart | josquin-dung | 1 | 505 | 4 |
|  |  | josquin-virgo | 1 | 950 | 4 |
|  |  | josquin-virgo | 2 | 1,047 | 4 |
| take15 | no-touch | josquin-dung | 1 | 505 | 4 |
|  |  | josquin-virgo | 1 | 1,044 | 4 |
|  |  | josquin-virgo | 2 | 1,038 | 4 |
| take16 | far-apart | josquin-dung | 1 | 505 | 4 |
|  |  | josquin-virgo | 1 | 1,050 | 4 |
|  |  | josquin-virgo | 2 | 1,046 | 4 |
| take17 | touch | josquin-dung | 1 | 506 | 4 |
|  |  | josquin-virgo | 1 | 1,052 | 4 |
|  |  | josquin-virgo | 2 | 1,051 | 4 |
| take18 | no-touch | dufay-gloria | 1 | 956 | 4 |
|  |  | dufay-kyrie1 | 1 | 792 | 4 |
|  |  | dufay-kyrie2 | 1 | 757 | 4 |
| take19 | far-apart | dufay-gloria | 1 | 970 | 4 |
|  |  | dufay-kyrie1 | 1 | 792 | 4 |
|  |  | dufay-kyrie2 | 1 | 761 | 4 |
| take20 | touch | dufay-gloria | 1 | 974 | 4 |
|  |  | dufay-kyrie1 | 1 | 770 | 4 |
|  |  | dufay-kyrie2 | 1 | 757 | 4 |
| take21 | no-touch | dufay-agnus1 | 1 | 625 | 4 |
|  |  | dufay-agnus2 | 1 | 432 | 2 |
|  |  | dufay-agnus3 | 1 | 607 | 4 |
|  |  | dufay-agnus1 | 2 | 625 | 4 |
|  |  | dufay-agnus2 | 2 | 434 | 2 |
| take22 | far-apart | dufay-agnus1 | 1 | 622 | 4 |
|  |  | dufay-agnus2 | 1 | 433 | 2 |
|  |  | dufay-agnus3 | 1 | 613 | 4 |
|  |  | dufay-agnus1 | 2 | 623 | 4 |
|  |  | dufay-agnus2 | 2 | 433 | 2 |
| take23 | touch | dufay-agnus1 | 1 | 624 | 4 |
|  |  | dufay-agnus2 | 1 | 435 | 2 |
|  |  | dufay-agnus3 | 1 | 610 | 4 |
|  |  | dufay-agnus1 | 2 | 625 | 4 |
|  |  | dufay-agnus2 | 2 | 434 | 2 |
| take28 | far-apart | dufay-gloria | 1 | 977 | 4 |
|  |  | dufay-kyrie1 | 1 | 777 | 4 |
|  |  | dufay-kyrie2 | 1 | 748 | 4 |
| take29 | touch | dufay-gloria | 1 | 968 | 4 |
|  |  | dufay-kyrie1 | 1 | 770 | 4 |
|  |  | dufay-kyrie2 | 1 | 752 | 4 |
| take30 | no-touch | dufay-gloria | 1 | 971 | 4 |
|  |  | dufay-kyrie1 | 1 | 789 | 4 |
|  |  | dufay-kyrie2 | 1 | 762 | 4 |
| take31 | far-apart | dufay-agnus1 | 1 | 619 | 4 |
|  |  | dufay-agnus2 | 1 | 431 | 2 |
|  |  | dufay-agnus3 | 1 | 613 | 4 |
|  |  | dufay-agnus1 | 2 | 615 | 4 |
|  |  | dufay-agnus2 | 2 | 434 | 2 |
| take32 | touch | dufay-agnus1 | 1 | 612 | 4 |
|  |  | dufay-agnus2 | 1 | 432 | 2 |
|  |  | dufay-agnus3 | 1 | 602 | 4 |
|  |  | dufay-agnus1 | 2 | 610 | 4 |
|  |  | dufay-agnus2 | 2 | 432 | 2 |
| take33 | no-touch | dufay-agnus1 | 1 | 615 | 4 |
|  |  | dufay-agnus2 | 1 | 431 | 2 |
|  |  | dufay-agnus3 | 1 | 602 | 4 |
|  |  | dufay-agnus1 | 2 | 615 | 4 |
|  |  | dufay-agnus2 | 2 | 432 | 2 |
| take37 | far-apart | josquin-dung | 1 | 503 | 4 |
|  |  | josquin-virgo | 1 | 1,042 | 4 |
|  |  | josquin-virgo | 2 | 1,048 | 4 |
| take38 | touch | josquin-dung | 1 | 506 | 4 |
|  |  | josquin-virgo | 1 | 1,047 | 4 |
|  |  | josquin-virgo | 2 | 1,047 | 4 |
| take39 | no-touch | josquin-dung | 1 | 502 | 4 |
|  |  | josquin-virgo | 1 | 1,048 | 4 |
|  |  | josquin-virgo | 2 | 1,050 | 4 |

Tab. A3. Independence (Permutation) tests for pitch performance indicators.

| Type | Indicator | Variable | Statistic | p |
| --- | --- | --- | --- | --- |
| Single voices | Mean Pitch Accurary | condition | 1.97 | 0.121 |
|  |  | day | 4.02 | 0.000\*\*\* |
|  |  | headset | 4.25 | 0.000\*\*\* |
|  |  | piece | 16.10 | 0.000\*\*\* |
|  | Mean Pitch Precision | condition | 3.57 | 0.001\*\* |
|  |  | day | 4.20 | 0.000\*\*\* |
|  |  | headset | 3.01 | 0.021\* |
|  |  | piece | 15.83 | 0.000\*\*\* |
| Within voices | Mean Pitch Accurary | condition | 3.87 | 0.000\*\*\* |
|  |  | day | 2.89 | 0.011\* |
|  |  | voice\_type | 5.11 | 0.000\*\*\* |
|  |  | piece | 8.61 | 0.000\*\*\* |
|  | Mean Pitch Precision | condition | 1.37 | 0.354 |
|  |  | day | 0.47 | 0.886 |
|  |  | voice\_type | 6.38 | 0.000\*\*\* |
|  |  | piece | 6.84 | 0.000\*\*\* |

Tab. A4. Independence (Permutation) tests for onset performance indicators.

| Type | Indicator | Variable | Statistic | p |
| --- | --- | --- | --- | --- |
| Between voices | Mean Onset Precision | condition | 1.13 | 0.497 |
|  |  | day | 1.04 | 0.553 |
|  |  | piece | 8.26 | 0.000\*\*\* |
| Within voices |  | condition | 0.74 | 0.738 |
|  |  | day | 1.03 | 0.559 |
|  |  | voice\_type | 8.58 | 0.000\*\*\* |
|  |  | piece | 8.07 | 0.000\*\*\* |