TOWARDS MEASURING INTONATION QUALITY OF CHOIR RECORDINGS: A CASE STUDY ON BRUCKNER'S LOCUS ISTE

Christof Weiß, Sebastian J. Schlecht, Sebastian Rosenzweig, Meinard Müller
International Audio Laboratories Erlangen, Germany
christof.weiss@audiolabs-erlangen.de

ABSTRACT

Unaccompanied vocal music is a central part of Western art music, yet it requires excellent skills for singers to achieve proper intonation. In this paper, we analyze intonation deficiencies by introducing an intonation cost measure that can be computed from choir recordings and may help to assess the singers' intonation quality. With our approach, we measure the deviation between the recording's local salient frequency content and an adaptive reference grid based on the equal-tempered scale. The adaptivity introduces invariance of the local intonation measure to global intonation drifts. In our experiments, we compute this measure for several recordings of Anton Bruckner's choir piece *Locus Iste*. We demonstrate the robustness of the proposed measure by comparing scenarios of different complexity regarding the availability of aligned scores and multi-track recordings, as well as the number of singers per part. Even without using score information, our cost measure shows interesting trends, thus indicating the potential of our method for real-world applications.

1. INTRODUCTION

Unaccompanied vocal music constitutes the nucleus of Western art music and the starting point of polyphony's evolution. Despite an increasing number of studies [1, 4, 5, 8–11, 16, 17, 21] dating back to the 1930s [29], many facets of polyphonic a cappella singing are yet to be explored and understood. A central challenge of a cappella singing is the adjustment of pitch in order to stay in tune relative to the fellow singers. Even if choirs achieve locally good intonation, they may suffer from intonation drifts slowly evolving over time [8–10, 15–17, 21, 23]. Thus, one has to deal with different intonation issues that refer to harmonic (or *vertical*) and melodic (or *horizontal*) intonation. In Fig. 1, we show how these different aspects of intonation quality may be visualized separately. Our schematic example illustrates the assessment of note-wise pitch deviations (color-coded) in the presence of a global pitch drift. Fig. 1a shows the idealized situation, where each note-wise pitch lies on a fixed reference grid. In Fig. 1b, all pitches are sharp (too high) compared to the same reference grid. The intonation quality, however, should be considered equivalent in both cases. Fig. 1c illustrates a different situation with downward intonation drift. Using a fixed reference grid, the deviations gradually accumulate. To compensate for this, one can use an adaptive reference grid [12, 15, 31] so that the vertical intonation quality is separated from the horizontal intonation drift. The residual pitch deviations—relative to the adaptive grid—only refer to vertical intonation problems (Fig. 1d), which are in the focus of our analysis.

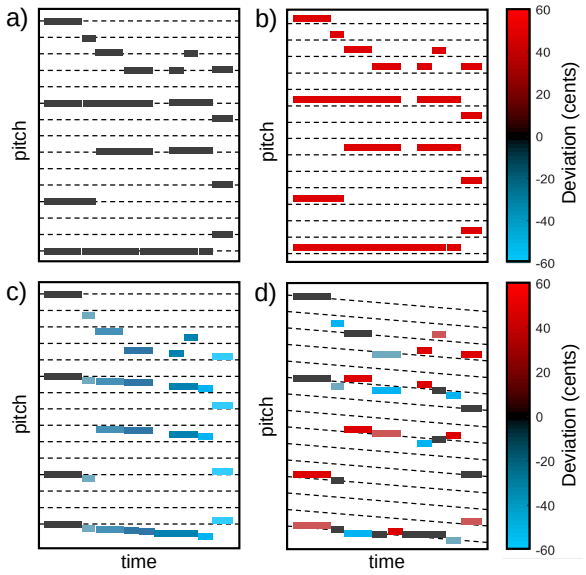

Figure 1. Note-wise analysis for polyphonic music. (a) Global intonation matching a fixed reference grid. (b) Global intonation higher than a fixed reference grid. (c) Intonation drift shown against a fixed reference grid. (d) Intonation drift with deviations from an adaptive grid.

In this paper, we propose an intonation cost measure that can be computed from choir recordings and may help to assess the singers' intonation quality. Developing such a measure encompasses two central challenges: (i) accurate estimation of the local salient frequency content from a choir recording, and, (ii) reliable measurement of intonation quality corresponding to human perception on the one hand and to music theory on the other hand.

Concerning the estimation of the local salient frequency content (i), recordings of polyphonic choir pieces constitute extremely difficult scenarios. Often, the different

parts of a musical composition are highly correlated both in rhythm (joint on- and offsets) and harmony (overlap of partials). In the case of a *mix* recording (single-track), this leads to overlaps in time–frequency representations, which makes the estimation of fundamental frequencies (F0s) [3] or partial tracking [25] hard. On the other hand, capturing intonation at the sub-semitone level requires high frequency resolutions. One can leverage these problems using dedicated recording scenarios where singers are isolated acoustically [5] or recorded sequentially [4]. Alternatively, special devices such as Larynx microphones [6, 16, 28] or additional score information [9] can help to simplify the F0-estimation problem [18, 27]. In Section 3.2, we detail the strategy used for this paper's experiments.

For assessing intonation quality (ii), we aim towards developing a robust intonation cost measure. Ideally, a high value of this measure indicates passages of low intonation quality. Hereby, intonation quality may relate to human perception such as the measure proposed in [30] based on psychometric curves [26]. On the other hand, intonation quality may be guided by music theory or historical performance practice [9]. In particular, choirs often aim for just intonation, which involves complex adjustment strategies according to the harmonic context [10]. In contrast to such ideas, we follow a simplified approach based on 12-tone equal temperament (12-TET). Even though 12-TET is not considered to be the ideal intonation practice for Western choir performance, it is a first approximation and can provide useful feedback [15]. As a major advantage, our strategy estimates the sub-semitone intonation quality for *any* type of notated chord irrespective of its *harmonic* consonance—in contrast to other methods [30] that measure a mixture of harmonic consonance and intonation.

As our main contribution, we propose a 12-TET-based intonation cost measure (Section 2). Inspired by prior work [15, 24], we accumulate the overall deviation of frequency components from an adaptive 12-TET grid weighted by their corresponding amplitudes. For testing this measure, we compiled a small but diverse dataset of Anton Bruckner's Graduale *Locus Iste* using different performances (Section 3). We evaluate the robustness of our method by comparing scenarios of varying complexity regarding the availability of aligned scores and multi-track recordings (Section 4.1). Finally, we apply our method to different performances and show its benefit for assessing the overall intonation quality of a recording (Section 4.2). Section 5 concludes the paper and gives an outlook on future work and practical applications.

2. MEASURING INTONATION QUALITY

In the following, we describe the computation of an intonation cost measure based on frequency deviations from a 12-tone equal-temperament (12-TET) grid.

2.1 Intonation Cost Measure

The proposed measure operates on a set of N frequency components $\mathcal{P} := \{(f_1, a_1), \ldots, (f_N, a_N)\}$, where each tuple $(f_n, a_n) \in \mathcal{P}$ denotes the frequency f_n and amplitude a_n of an individual component.

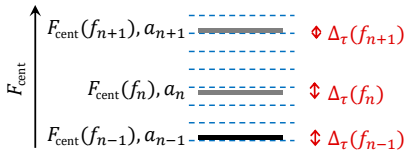

Figure 2. Grid deviations $\Delta_\tau(f_n)$ in cents of frequency components f_n (solid gray lines) from a 12-TET grid (dashed blue lines) shifted by τ. The corresponding amplitudes a_n are indicated by the grayscale colors.

First, we convert a given frequency f in Hertz (Hz) to cents by

$$F_{\text{cent}}(f) := 1200 \cdot \log_2\left(\frac{f}{f_0}\right), \tag{1}$$

where $f_0 := 55\,\text{Hz}$ is an arbitrary but fixed reference frequency. We compute the deviation of the frequency component f from a 12-TET grid

$$\Delta_\tau(f) := \min_{i \in \mathbb{Z}} |F_{\text{cent}}(f) - \tau - 100i|, \tag{2}$$

where $\tau \in [-50, 50[$ specifies the overall grid shift in cents, see Fig. 2. Applying a Gaussian-like function to the grid deviation $\Delta_\tau(f)$, we define the intonation cost (IC) Θ_τ as

$$\Theta_\tau(\mathcal{P}) := \frac{\sum_{(f,a)\in\mathcal{P}} a\left(1 - \exp\left(-\frac{\Delta_\tau^2(f)}{2\sigma^2}\right)\right)}{\sum_{(f,a)\in\mathcal{P}} a}, \tag{3}$$

where the deviations are weighted and then normalized by the corresponding amplitudes. Due to the normalization, we obtain $\Theta_\tau(\mathcal{P}) \in [0, 1]$ where $\Theta_\tau(\mathcal{P}) = 1$ indicates the maximal IC. The parameter σ adjusts the cost for deviating from the grid. As suggested by [24], we choose a value of $\sigma = 16$ cents. To obtain invariance to pitch drifts, i.e., variation of the reference frequency, we choose the grid shift τ in an adaptive way so that the IC is minimized:

$$\tau^* := \operatorname*{arg\,min}_{\tau \in [-50,50[} \Theta_\tau(\mathcal{P}). \tag{4}$$

We then define the intonation cost Θ as

$$\Theta(\mathcal{P}) := \Theta_{\tau^*}(\mathcal{P}). \tag{5}$$

For instance, in a scenario where a choir performs with good local intonation but is affected by a pitch drift, τ^* slowly changes over time while Θ is constantly small.

2.2 Example with Synthetic Signals

In the following, we illustrate the properties of the IC measure $\Theta(\mathcal{P})$ by means of synthetic examples. To this end, we define a harmonic tone $x_f : \mathbb{R} \to \mathbb{R}$ with K partials and fundamental frequency f as

$$x_f(t) := \sum_{k=1}^{K} a_k \cdot \sin(2\pi k f t), \tag{6}$$

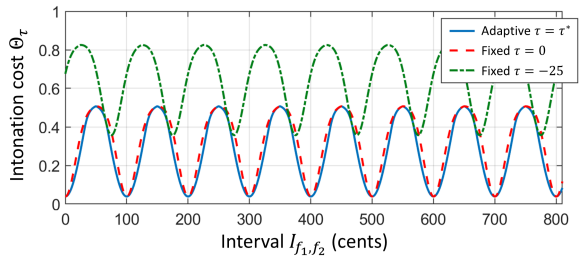

Figure 3. ICs Θ_τ for two harmonic tones x_{duad} with an interval of size I_{f_1,f_2} in cents and three grid shifts: adaptive grid τ^*, fixed grid $\tau = 0$, and, fixed grid $\tau = -25$. The F0 of the lower tone is fixed at $f_1 = 220$ Hz.

where t denotes time and

$$a_k := s^{k-1} \tag{7}$$

denotes the geometrically decaying partial amplitudes for some $s \in [0, 1]$ and $k = 1, \ldots, K$. Thus, the set of frequency components of signal x_f is

$$\mathcal{P}[x_f] := \{(f, a_1), (2f, a_2), \ldots, (Kf, a_K)\}. \tag{8}$$

Let $x_{\text{duad}} := x_{f_1} + x_{f_2}$ be the sum of two harmonic tones whose fundamental frequencies differ by the interval

$$I_{f_1,f_2} := |F_{\text{cent}}(f_2) - F_{\text{cent}}(f_1)| \tag{9}$$

given in cents. Consequently, $\mathcal{P}[x_{\text{duad}}] = \mathcal{P}[x_{f_1}] \cup \mathcal{P}[x_{f_2}]$.

Fig. 3 shows $\Theta(\mathcal{P}[x_{\text{duad}}])$ for two harmonic tones with $K = 16$ and $s = 0.6$ for different interval sizes I_{f_1,f_2}. The lower fundamental frequency is set to $f_1 = 220$ Hz such that $\Delta_0(f_1) = 0$. $\Theta(\mathcal{P}[x_{\text{duad}}],)$ is minimal for I_{f_1,f_2} being an integer multiple of 100 cents. However, $\Theta(\mathcal{P}[x_{\text{duad}}])$ does not reach zero as some partial frequencies $k \cdot f$ of a harmonic tone do not lie on the 12-TET grid even if the fundamental frequency f does. For example, the third partial $3f_1 = 660$ Hz leads to $\Delta_0(3f_1) \approx 2$ cents and the fifth partial $5f_1 = 1100$ Hz leads to $\Delta_0(5f_1) \approx 14$ cents. Since the minimal values are close to zero, this effect is small for a partial decay of $s = 0.6$. On the other hand, even a quarter tone interval $I_{f_1,f_2} = 50$ cents does not lead to the maximal IC of 1, since the grid deviation $\Delta_{\tau^*}(k \cdot f_1) \approx \Delta_{\tau^*}(k \cdot f_2) \approx 25$ cents for $\tau^* = 25$ cents. Fig. 3 further shows that the IC of a fixed grid shift $\tau = 0$ is similar to the adaptive grid τ^*, while a fixed shift $\tau = -25$ significantly increases the overall IC. It is important to note that the IC with adaptive grid shift τ^* is invariant to the choice of f_1 while a fixed grid shift is not. Further, the minimal and maximal values depend on the amplitude decay s and on the Gaussian width σ. Since the IC measure relates to a 12-TET grid, the IC curve is periodic in interval size with a period of 100 cents. As a consequence, each musical interval is only evaluated by its deviation from the 12-TET scale—regardless of its harmonic consonance quality.

In Fig. 4, we expand the previous example to three harmonic tones $x_{\text{triad}} := x_{f_1} + x_{f_2} + x_{f_3}$ with $f_1 \leq f_2 \leq f_3$. For instance, the intervals $I_{f_1,f_2} = 400$ cents, $I_{f_2,f_3} = 300$ cents describe an equal-tempered major triad. The colors in Fig. 4 indicate $\Theta(\mathcal{P}[x_{\text{triad}}])$ with respect to the

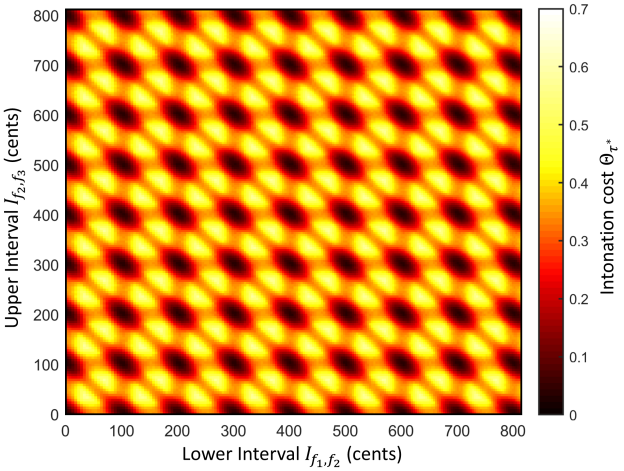

Figure 4. IC Θ for three harmonic tones x_{triad}. The plot axes indicate the size of the lower and upper interval in cents, I_{f_1,f_2} and I_{f_2,f_3} respectively.

lower and upper intervals, I_{f_1,f_2} and I_{f_2,f_3}, respectively. Similar to Fig. 3, we observe a periodic structure with period 100 cents as the IC is invariant to the musical intervals. Thus, the measure is equally suited for estimating the intonation quality of both consonant and dissonant triads.

3. EXPERIMENTAL SCENARIO

This section describes the experimental scenario for applying our intonation cost measure to choir recordings.

3.1 Dataset

We compiled a small but diverse dataset of performances of Anton Bruckner's Gradual *Locus iste* WAB 23 (see Fig. 5). This choir piece is in Latin and lasts approximately three minutes. It is musically interesting, contains several melodic and harmonic challenges—such as the highly chromatic middle part—but also harmonically clear passages, and covers a large part of each voice's tessitura.

Central to this dataset is a publicly available [1] multitrack recording from the Choral Singing Dataset (CSD) [4]. The performance of 16 singers from the semi-professional Anton Bruckner choir (Barcelona) was recorded in a studio setting. The four musical parts—soprano, alto, tenor, and bass—were recorded sequentially using directional hand-held microphones. Rhythmic and harmonic synchronization was ensured by a conducting video and an acoustic reference (MIDI version of the piece). Due to this recording scenario, the individual singers' tracks exhibit a small amount of bleeding from other singers of the same part (e. g., soprano 2, 3, and, 4 slightly bleed into soprano 1 track). Interactive intonation or adaptation across musical parts was not possible since the parts were recorded in isolation and each singer listened to the reference MIDI signal while singing—this also prevented substantial pitch drifts. The restricted interaction limits the usability of the data to study intonation and

[1] https://zenodo.org/record/1319597#.XJor8ShKhaR

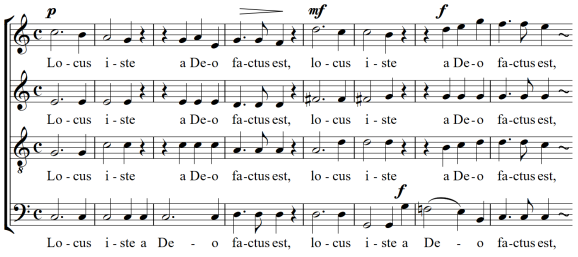

Figure 5. Anton Bruckner, *Locus iste* WAB 23, beginning.

adaptation phenomena in choir performances. Nevertheless, since this paper focuses on measurement strategies, the multi-track data provides an excellent resource. To address subsets of the multi-track recording, we refer to the signals of the four soprano voices as x_{S1}, x_{S2}, x_{S3}, and x_{S4}, and to the first voices of the alto, tenor, and bass as x_{A1}, x_{T1}, and x_{B1}, respectively. We denote the mixed signal of the first voices of each part as x_{SATB1} and the mixed signal of all 16 voices as x_{SATB}. We further manipulated the original tracks x^{orig} with the digital audio correction software Melodyne [2] to augment the dataset: x^{note} is generated by quantizing the median pitch of each note event onto a 12-TET grid with reference frequency $f_0 = 55$ Hz. x^{fine} is generated by quantizing the complete F0-trajectory onto the 12-TET grid. To assist our analysis, we use additional score information from the aligned MIDI file.

In addition to this multi-track recording, we collected several commercial [3] and freely available [4] performances of the piece. As a reference for the real audio measurements, we sonified the piece with harmonic tones (Section 2.2) using random pitch deviations sampled from Gaussian distributions with different standard deviations.

3.2 Measuring Frequency Content

As discussed in Section 1, the extraction of salient frequency content from choir recordings is challenging. In the case of a mix recording, we have to blindly estimate all partial frequencies using a partial tracking algorithm such as [25]. For choir music, partial tracking can be simplified as the singing voice's partial frequencies are located quite precisely at integer multiples of an estimated F0. To estimate F0-trajectories for the CSD, we can use the individual tracks of the multi-track recording. Due to the bleeding of other voices from the same part, traditional F0-estimation techniques [7, 20] may have problems. Therefore, we use a salience-based method similar to Melodia [27]. We compute a log-frequency representation using instantaneous frequency estimation [2, 14, 27] and binning with a resolution of $1\,\mathrm{cent}$. Subsequently, we estimate F0-trajectories using dynamic programming [22]. As post-processing, we apply median filtering with a filter length of 101 frames and downsample by a factor of 50 obtaining trajectories with a time resolution of $290\,\mathrm{ms}$.

[2] `https://www.celemony.com`

[3] Philharmonia Vocalensemble Stuttgart (Profil Medien 2006), Chor des Bayerischen Rundfunks (Decca 2012), Choir of St John's College Cambridge (Classic Mania 2007), NDR Chor Hamburg (Carus 2015)

[4] Internet Archive, `https://archive.org/details/LocusIste`

In the case of the CSD, we can exploit additional score information from the aligned MIDI file [13]. We restrict the F0-estimation to rectangular time–frequency regions ("constraint regions") derived from onsets, durations and center frequencies of the aligned MIDI notes, including a frequency tolerance of $\pm 60\,\mathrm{cents}$ around the center frequencies (according to 12-TET with reference 440 Hz). Such additional information is particularly helpful when estimating F0-trajectories from a mix recording. [5] Especially, the constraint regions prevent the common confusion of the F0 with higher partials.

4. RESULTS

In our experiments, we investigate the robustness of the proposed intonation measure for different scenarios of the CSD where either score information or multi-track recordings are not used (Section 4.1). Furthermore, we compare the measure's behaviour for different synthetic and real performances of *Locus Iste* (Section 4.2).

4.1 Local Analysis and Visualization

As a visual orientation, we show in Fig. 6a a piano roll representation of the entire piece, generated from the aligned MIDI file. Considering the subset of each part's first voice x_{S1}, x_{A1}, x_{T1}, and x_{B1}, we estimate F0-trajectories as described in Section 3.2 and compute the local, sub-semitone deviation of the F0-estimate from the corresponding MIDI reference. The deviations are color-coded with a range of $\pm 60\,\mathrm{cents}$. While there seems to be no significant global drift (which is not surprising because of the CSD's recording scenario, see Section 3.1), we observe a slight dominance of notes sung flat (negative deviation) except for the alto part, which is sharp more often.

The main results are shown in Figures 6b–g, which indicate IC values throughout each performance. We compute the IC measure $\Theta(\mathcal{P})$ of the set $\mathcal{P}$ comprising the frequencies (F0 and higher partials) and amplitudes from all parts that are, according to the aligned score, active in a frame. Assuming that the human voice's partials are harmonic, we calculate the first 16 partial frequencies from the measured F0-trajectories and extract the corresponding amplitudes from the log-frequency spectrogram. For computing Θ, we use the adaptive grid shift proposed in Section 2.1. To remove local outliers, we post-process the IC curves using a moving median filter with a length of 21 frames.

The blue, solid line in Fig. 6b shows the resulting IC curve for x^{orig}, computed from the individual tracks for the first voice of each part x_{S1}, x_{A1}, x_{T1}, and x_{B1} using score constraints. For silent regions (e. g., after 160 sec), the IC is zero since no constraint region is active. Due to the adaptive grid shift, the IC is small for monophonic passages where only one singer is active (see, e. g., the passage at 80 sec). For some of the consonant chords (e. g., at 110 sec), we observe low IC values of about 0.2. During the highly chromatic three-part passage (80–110 sec), the

[5] We do not use harmonic summation as in [27] to avoid smearing of other parts' partials into the constraint regions for mix recordings.

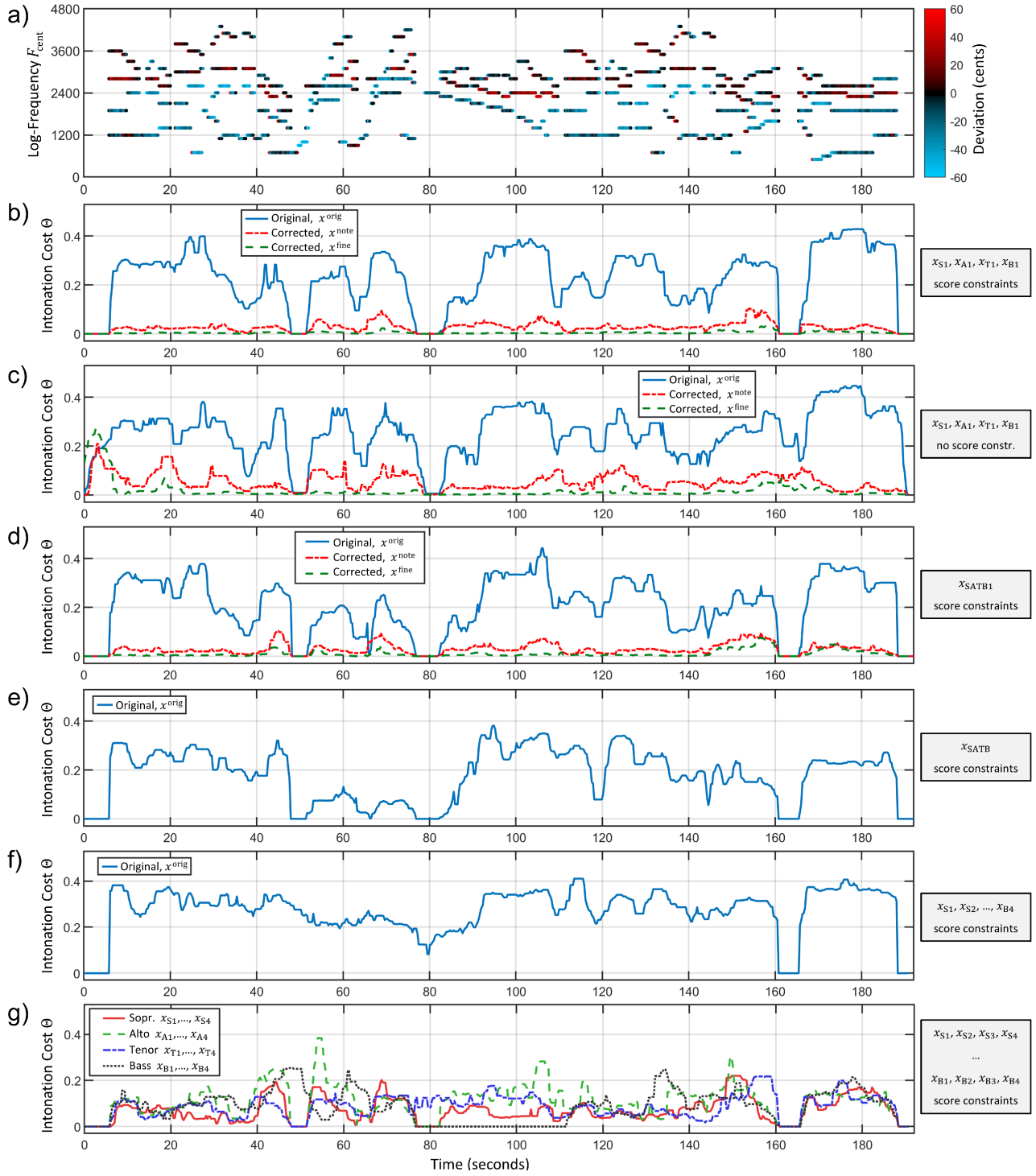

Figure 6. Intonation cost (IC) curves for different CSD versions of *Locus Iste*. (a) Piano roll representation with a log-frequency axis where C2 corresponds to 1200 cents, C3 to 2400 cents etc. The colors encode the deviation of the first voice of each part (x_{S1}, x_{A1}, x_{T1}, x_{B1}) from the MIDI reference. (b) IC curves for x_{S1}, x_{A1}, x_{T1}, x_{B1} with score constraints from four individual tracks. The blue curve corresponds to x^{orig}, the red curve to x^{note}, and the green dashed curve to x^{fine}. (c) IC curves as in (b) without score constraints. (d) IC curves for mixed signal x_{SATB1} with score constraints. (e) IC curve for mixed signal of all voices x_{SATB} with score constraints. (f) IC curve computed from all 16 individual tracks x_{S1}, x_{S2}, ..., x_{B4} with score constraints. (g) Individual IC curves for the four tracks of each part with score constraints.

IC increases, thus indicating that the singers have difficulties to stay in tune in this passage.

As a sanity check, we compare the results for x^{orig} to the the pitch-corrected versions x^{note} and x^{fine}. As we expected, the note-wise corrections x^{note} (red curve) obtain lower values than x^{orig}—with only minor peaks that may be caused by pitch fluctuations during a note event. If we correct such local fluctuations as in x^{fine} (green, dashed curve), the IC is almost constantly zero.

In Fig. 6c, we repeat the experiment of Fig. 6b without using score constraints. In this case, F0-estimation is less reliable and, in particular, confusions between F0 and higher partials may occur. However, since many partials lie on the same 12-TET grid as the F0 (octave-related partials) or very close to that (2 cents for fifth-related partials), the IC measure is largely invariant to such confusions. The high similarity between Fig. 6b and Fig. 6c confirms the IC measure's robustness to F0-extraction errors. Only for silent passages (e. g., after 160 sec or at the beginning), we obtain higher IC values due to erroneously extracted frequency components. [6] We conclude that our strategy is, in principle, applicable for any multi-track recording and does not necessarily require score information.

To test the applicability in absence of multi-track recordings, we compute the IC curve from the mix signal x_{SATB1} using score constraints (Fig. 6d). Due to the constraints, confusions with higher partials cannot occur, but partials of other parts may leak into the same constraint regions, thus affecting the estimated F0s and partials' amplitudes. The comparison of Fig. 6b and Fig. 6d indicates that such phenomena only slightly affect the IC measure.

Next, we measure the IC from the CSD's full recording x_{SATB} (all 16 voices). F0-estimation is more challenging since the four singers of each part contribute to the same constraint regions. The resulting ICs (Fig. 6e) exhibit slightly lower values than in previous cases. We assume that having several voices per part stabilizes the F0-estimation to some degree. Overall, we see a similar trend between Fig. 6d and Fig. 6e. This might be an effect of mutual influence between the singers of each part. To further investigate the multiple-singer effect, we show in Fig. 6f the IC curves computed from all 16 individual tracks $x_{\mathrm{S1}}, x_{\mathrm{S2}}, \ldots, x_{\mathrm{B4}}$. We obtain higher IC values than in Fig. 6e, especially for the monophonic passages (e. g., at 48 sec). Due to the score constraints, this must be caused by deviations between the singers of a part, sometimes denoted as *dispersion* [4]. To analyze this, we repeat the experiment for each part separately (Fig. 6g). The resulting curve supports our hypothesis since. For instance, the high value at around 48 sec (Fig. 6f) is mainly caused by the basses' dispersion (dotted curve in Fig. 6g).

4.2 Global Analysis of Different Performances

Since our experiments on the CSD have shown the robustness of our method even for mix recordings, we finally compare the global IC values of multiple synthetic and real performances of *Locus Iste*. We align the MIDI file to all performances [13], use the resulting constraint regions for extracting F0-trajectories [27], and compute the IC curves. In Fig. 7, we show the statistics of the entire curves over time (median, mean, and standard deviation). First, we report values for sonifications using harmonic tones as defined in Eq. (6). To simulate detuning, we shifted each note's fundamental frequency by a random value sampled from a Gaussian distribution with variance

[6] This problem could be leveraged using a suitable algorithm for voice activity detection [19] or on-/offset detection in vocal music [6].

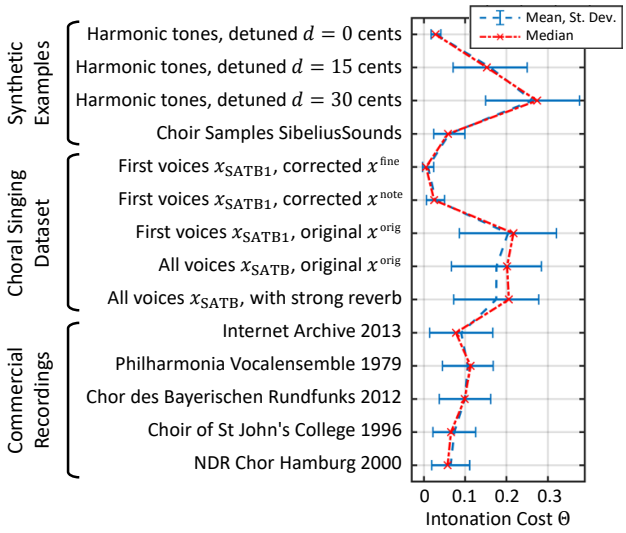

Figure 7. Statistics of the intonation cost measure Θ over full recordings of different type.

d^2. For $d = 0$ cents, the mean IC is almost zero. For $d = 15$ cents and $d = 30$ cents, the IC gradually increases as expected. The average IC of a sample-based sonification [7] is moderately higher than the rendition with ideal harmonic tones, which suggests that some choir effects such as dispersion are also synthesized. Both corrected versions of the CSD show very low IC values, with x^{fine} even lower than x^{note}. In contrast, the versions based on the original recording x^{orig} have high IC values, where the difference between first voices only (x_{SATB1}) and the full mix (x_{SATB}) as well as the effect of adding artificial reverberation is marginal. All commercial recordings performed by professional choirs obtain lower IC than the CSD recording. Several effects might contribute to this observation. Besides the singers' level of training, commercial recordings often feature larger choirs (60 singers or more) and long natural reverberation (recorded in churches). In line with the authors' subjective judgment, the recording by NDR Chor Hamburg exhibits the best intonation quality according to the IC measure. [8] Overall, this experiment indicates that the proposed IC measure may serve as a first indicator for a choir recording's global intonation quality.

5. CONCLUSIONS

In this paper, we proposed a strategy for measuring the intonation quality of choir recordings. Although robust extraction of salient frequencies is challenging, the measure produced meaningful and reliable results once multi-track recordings are available or score information can be utilized. Even though the insights from such a measure are limited, it might be a first indicator for the overall intonation quality and, thus, could be useful for choir singers or choir directors in performances and rehearsals.

[7] Using Sibelius Sounds, see `https://www.avid.com/sibelius`

[8] A 30-seconds mp3 thumbnail of this recording is available at `https://www.carusmedia.com/images-intern/medien/80/8346600/8346600.010s.t1_010.mp3`

Acknowledgments: This work was supported by the German Research Foundation (DFG MU 2686/12-1). The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institut für Integrierte Schaltungen IIS. The authors want to thank Helena Cuesta and colleagues from UPF Barcelona for creating and publishing the Choral Singing Dataset.

6. REFERENCES

[1] Per-Gunnar Alldahl. *Choral Intonation*. Gehrmans Musikförlag, 1990.

[2] François Auger and Patrick Flandrin. Improving the readability of time-frequency and time-scale representations by the reassignment method. *IEEE Transactions on Signal Processing*, 43(5):1068–1089, 1995.

[3] Mert Bay, Andreas F. Ehmann, and J. Stephen Downie. Evaluation of multiple-f0 estimation and tracking systems. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 315–320, Kobe, Japan, 2009.

[4] Helena Cuesta, Emilia Gómez, Agustín Martorell, and Felipe Loáiciga. Analysis of intonation in unison choir singing. In *Proceedings of the International Conference of Music Perception and Cognition (ICMPC)*, pages 125–130, Graz, Austria, 2018.

[5] Jiajie Dai and Simon Dixon. Analysis of interactive intonation in unaccompanied SATB ensembles. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 599–605, Suzhou, China, 2017.

[6] Sara D'Amario, Helena Daffern, and Freya Bailes. A new method of onset and offset detection in ensemble singing. *Logopedics Phoniatrics Vocology*, 2018.

[7] Alain de Cheveigné and Hideki Kawahara. YIN, a fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America (JASA)*, 111(4):1917–1930, 2002.

[8] Johanna Devaney. *An Empirical Study of the Influence of Musical Context on Intonation Practices in Solo Singers and SATB Ensembles*. PhD thesis, McGill University, Montreal, Canada, 2011.

[9] Johanna Devaney and Daniel P. W. Ellis. An empirical approach to studying intonation tendencies in polyphonic vocal performances. *Journal of Interdisciplinary Music Studies*.

[10] Johanna Devaney, Michael I. Mandel, and Ichiro Fujinaga. A study of intonation in three-part singing using the automatic music performance analysis and comparison toolkit (AMPACT). In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 511–516, Porto, Portugal, 2012.

[11] Johanna Devaney, Jonathan Wild, and Ichiro Fujinaga. Intonation in solo vocal performance: A study of semitone and whole tone tuning in undergraduate and professional sopranos. In *Proceedings of the International Symposium on Performance Science*, pages 219–224, Toronto, Canada, 2011.

[12] Karin Dressler and Sebastian Streich. Tuning frequency estimation using circular statistics. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 357–360, Vienna, Austria, 2007.

[13] Sebastian Ewert, Meinard Müller, and Peter Grosche. High resolution audio synchronization using chroma onset features. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1869–1872, Taipei, Taiwan, April 2009.

[14] J. L. Flanagan and R. M. Golden. Phase vocoder. *Bell System Technical Journal*, 45:1493–1509, 1966.

[15] Volker Gnann, Markus Kitza, Julian Becker, and Martin Spiertz. Least-squares local tuning frequency estimation for choir music. In *Proceedings of the Audio Engineering Society (AES) Convention*, New York City, USA, 2011.

[16] David M. Howard. Intonation drift in a capella soprano, alto, tenor, bass quartet singing with key modulation. *Journal of Voice*, 21(3):300–315, 2007.

[17] David M. Howard, Helena Daffern, and Jude Brereton. Four-part choral synthesis system for investigating intonation in a cappella choral singing. *Logopedics Phoniatrics Vocology*, 38(3):135–142, 2013.

[18] Jong Wook Kim, Justin Salamon, Peter Li, and Juan Pablo Bello. Crepe: A convolutional representation for pitch estimation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 161–165, Calgary, Canada, 2018.

[19] Bernhard Lehner, Jan Schlüter, and Gerhard Widmer. Online, loudness-invariant vocal detection in mixed music signals. *IEEE/ACM Transactions on Audio, Speech & Language Processing*, 26(8):1369–1380, 2018.

[20] Matthias Mauch and Simon Dixon. pYIN: A fundamental frequency estimator using probabilistic threshold distributions. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 659–663, Florence, Italy, 2014.

[21] Matthias Mauch, Klaus Frieler, and Simon Dixon. Intonation in unaccompanied singing: Accuracy, drift, and a model of reference pitch memory. *Journal of the Acoustical Society of America*, 136(1):401–411, 2014.

[22] Meinard Müller. *Fundamentals of Music Processing*. Springer Verlag, 2015.

[23] Meinard Müller, Peter Grosche, and Frans Wiering. Automated analysis of performance variations in folk song recordings. In *Proceedings of the International Conference on Multimedia Information Retrieval (MIR)*, pages 247–256, Philadelphia, Pennsylvania, USA, 2010.

[24] Tomoyasu Nakano, Masataka Goto, and Yuzuru Hiraga. An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 1706–1709, Pittsburgh, PA, USA, 2006.

[25] Julian Neri and Philippe Depalle. Fast partial tracking of audio with real-time capability through linear programming. In *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, pages 326–333, Aveiro, Portugal, 2018.

[26] Reinier Plomp and Willem Johannes Maria Levelt. Tonal consonance and critical bandwidth. *Journal of the Acoustical Society of America*, 38(4):548–560, 1965.

[27] Justin Salamon and Emilia Gómez. Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(6):1759–1770, 2012.

[28] Frank Scherbaum. On the benefit of larynx-microphone field recordings for the documentation and analysis of polyphonic vocal music. *Proceedings of the International Workshop Folk Music Analysis*, pages 80–87, 2016.

[29] Carl Emil Seashore. *Objective Analysis of Musical Performance*, volume 4 of *Studies in the Psychology of Music*. University of Iowa Press, Iowa City, USA, 1936.

[30] William A. Sethares. Local consonance and the relationship between timbre and scale. *Journal of the Acoustical Society of America*, 94(3):1218–1228, 1993.

[31] Simon Waloschek and Aristotelis Hadjakos. Driftin' down the scale: Dynamic time warping in the presence of pitch drift and transpositions. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, pages 630–636, Paris, France, 2018.