A DIFFERENTIABLE COST MEASURE FOR INTONATION PROCESSING IN POLYPHONIC MUSIC

Simon Schwär **Sebastian Rosenzweig** **Meinard Müller**
International Audio Laboratories Erlangen, Germany
simon.schwaer@fau.de, {sebastian.rosenzweig, meinard.mueller}@audiolabs-erlangen.de

ABSTRACT

Intonation is the process of choosing an appropriate pitch for a given note in a musical performance. Particularly in polyphonic singing, where all musicians can continuously adapt their pitch, this leads to complex interactions. To achieve an overall balanced sound, the musicians dynamically adjust their intonation considering musical, perceptual, and acoustical aspects. When adapting the intonation in a recorded performance, a sound engineer may have to individually fine-tune the pitches of all voices to account for these aspects in a similar way. In this paper, we formulate intonation adaptation as a cost minimization problem. As our main contribution, we introduce a differentiable cost measure by adapting and combining existing principles for measuring intonation. In particular, our measure consists of two terms, representing a tonal aspect (the proximity to a tonal grid) and a harmonic aspect (the perceptual dissonance between salient frequencies). We show that, combining these two aspects, our measure can be used to flexibly account for different artistic intents while allowing for robust and joint processing of multiple voices in real-time. In an experiment, we demonstrate the potential of our approach for the task of intonation adaptation of amateur choral music using recordings from a publicly available multitrack dataset.

1. INTRODUCTION

The widely-used 12-tone equal temperament (12-TET) tuning system divides the octave in twelve equal semitones of the ratio $2^{1/12} \approx 1.0595$. This allows instruments with fixed pitch to play in any key at the cost of most intervals being slightly out of tune in comparison to the natural overtone spectrum of harmonic sounds. Just Intonation (JI) scales, on the other hand, are constructed from intervals with small integer ratios to a root note. As a result, the harmonic overtones of two tones in a JI scale are more congruent than those in 12-TET. However, the absolute pitches in the JI scale change for different root notes, so that the grid must be adapted to different keys and musical contexts.

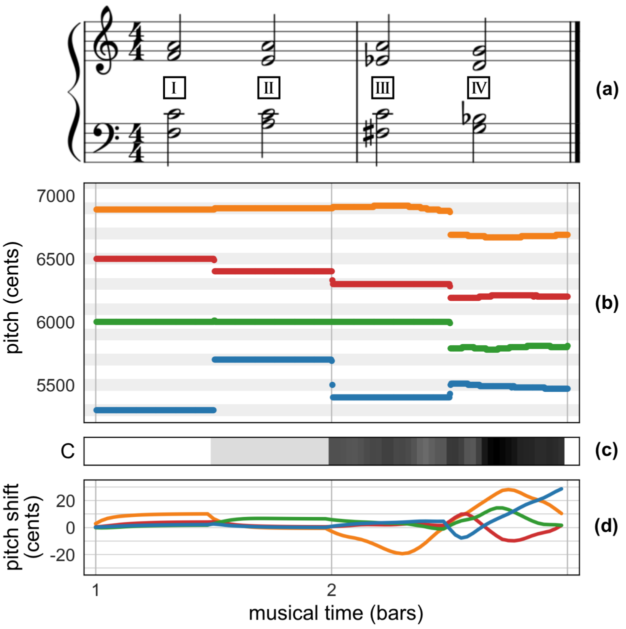

Figure 1: Joint adaptation of the voices in an example cadence. **(a)** Sheet music. **(b)** Fundamental frequencies of the original synthesized voices measured with pYIN [1] (orange: soprano, red: alto, green: tenor, blue: bass). **(c)** Overall intonation cost C between all voices with $w = 0.33$. **(d)** Pitch shift curve for all voices obtained from joint gradient descent on C with $w = 0.33$ and $\mu = 350$.

Many instruments can produce any pitch in between the 12-TET or JI grid or have considerable variance in tuning. This allows performers to dynamically change the sounding scale or chord, both intentionally and accidentally. This flexible intonation is particularly relevant in a cappella choral singing. While 12-TET if often used as an approximation for the distribution of chosen pitches by singers [2, 3], their intonation is influenced by a multitude of aspects. For example, choir singers tend to aim for JI in harmonies [4], whereas other influences may prevail in melodic or solistic phrases [5]. At the same time, singers continuously have to account for the intonation of their fellow musicians [6, 7], while pitch changes also occur during the sounding tone [8]. Depending on the singers' ability to control their voices, this complex setting often results in defects like poor local intonation or intonation drift [9, 10].

Different aspects of intonation are illustrated in the synthesized example cadence shown in Figure 1: The pitches

Attribution: S. Schwär, S. Rosenzweig, and M. Müller, "A Differentiable Cost Measure for Intonation Processing in Polyphonic Music", in *Proc. of the 22nd Int. Society for Music Information Retrieval Conf.*, Online, 2021.

(where "pitch" is used as a technical term synonymously with fundamental frequency here and in the following) in chord I correspond to a JI scale with root F, while chord II is tuned to 12-TET. Continuous deviations are illustrated in chords III and IV, where the soprano in III and all voices in IV are detuned randomly between -25 and $+25$ cents around the 12-TET grid. Even with these large pitch fluctuations, the chords can still be clearly recognized [1] .

When post-processing multitrack recordings, pitch-shifting the individual audio signals may mitigate some unintended intonation deviations in a performance. However, this requires a known target pitch, and similarly to intonation in a live performance, the desired target can be influenced by many aspects. Instead of quantizing to a fixed set of pitches like 12-TET or manually tweaking individual notes, we formulate intonation adaptation as a cost minimization problem. A good cost measure for this task should have a local minimum at the target pitch for each individual note.

As our main contribution, we propose a differentiable cost measure, where the local minima can be adjusted according to artistic intent. In particular, we employ two existing models to account for different aspects of intonation: The first model represents a *tonal* aspect, that most music is composed from a set of discrete pitches approximated by equal divisions of the octave [11]. The second model considers a *harmonic* aspect and uses perceptual dissonance to capture the tendency for JI in multi-part harmonies [12].

We show that our cost-based approach has several advantageous properties over existing methods for intonation processing:

- A variable weight between the terms allows for flexibly setting the local minimum anywhere between 12-TET and JI.
- In contrast to purely dissonance-based adaptation, our cost measure has a local minimum also for musically unstable voices of a chord.
- Using gradient descent, intonation can be adapted in real-time, dynamically reacting to changing inputs.

For example, the overall cost shown in Figure 1c is high when voices deviate strongly from the desired pitches. At the same time, musically dissonant chords like the diminished seventh chord in III have a higher inherent perceptual dissonance. Therefore, an adaptation should not aim to achieve zero cost, but to find the nearest local minimum.

Figure 1d shows the pitch shift curves for the voices in our example that locally minimize the cost measure. The curves were obtained using joint gradient descent, where all voices are processed at the same time and influence each other. The resulting "optimal" pitches after applying the shift lie in between 12-TET and JI, as can be seen e. g. in the major third of chord I. Its initial pitch in the present example is -14 cents w.r.t. 12-TET and it is pitched up by 10 cents to minimize the cost with the given parameters. Furthermore, the algorithm finds meaningful solutions in the more complex situations occuring in chords III and IV.

The remainder of this article is structured as follows. In Section 2, we review existing approaches to intonation adaptation and adaptive tuning, in Section 3 we introduce the cost measure, and in Section 4, we demonstrate the applicability to local intonation adaptation in a multitrack choral music recording with amateur singers.

2. RELATED WORK

A common intonation adaptation strategy implemented in many commercial products like *Melodyne* [13] or *Auto-Tune* [14] is to measure the fundamental frequency (F0) in a monophonic recording and to pitch-shift the signal such that the F0 approaches a fixed target value. The target can be chosen manually by the user or determined automatically from a predefined grid or score.

Several approaches have been proposed to dynamically choose a target pitch based on musical assumptions. Rule-based algorithms like *Groven.Max* [15] or *Hermode Tuning* [16] choose pitches for all voices of a synthesizer by analyzing the musical structure of a chord. Aiming for JI, they implement fixed rules to compromise in chords where just intervals between all pairs of notes are not possible. This problem can also be addressed by solving a quadratic program [17]. This way, the deviation from JI is distributed evenly across the pitches and all intervals are as close as possible to a small integer ratio. Additional constraints can enable temporal continuity.

Deep learning is used in [18] to infer "good" intonation from curated training examples of monophonic vocal recordings over a backing track. The model then outputs a pitch shift curve that can match the intonation in an input recording with the characteristics of the training examples.

Sethares [19] relates the chosen scale to the timbre of the sound. Summing the perceptual dissonance [20] of all individual salient frequency pairs between two sounds, he obtains a *dissonance landscape*, in which local minima exist for small integer ratio intervals if the timbre is harmonic. This principle is also used for adaptive tuning using gradient descent [12], which achieves a tuning similar to JI without requiring explicit musical analysis of the chords. In [21], the idea was further enhanced to be stable in more complex settings by adding a proximity constraint. This limits the deviation from 12-TET to a few cents and requires the input to be in the same range.

3. INTONATION COST MEASURE

Our cost measure is based on the assumption that proper intonation in a polyphonic context is a balance between the proximity to pitches in an equal temperament tuning system most suitable for the composition and the minimization of perceptual dissonance. Mathematically, this can be expressed as the sum of a *tonal* cost C_t, indicating the distance to the pitches in an equal temperament tuning system, and a *harmonic* cost C_h, measuring the perceptual dissonance in the overall sound:

$$C := w \cdot C_\mathrm{t} + (1 - w) \cdot C_\mathrm{h}. \tag{1}$$

[1] Audio examples are available online: `https://www.audiolabs-erlangen.de/resources/MIR/2021-ISMIR-IntonationCostMeasure`.

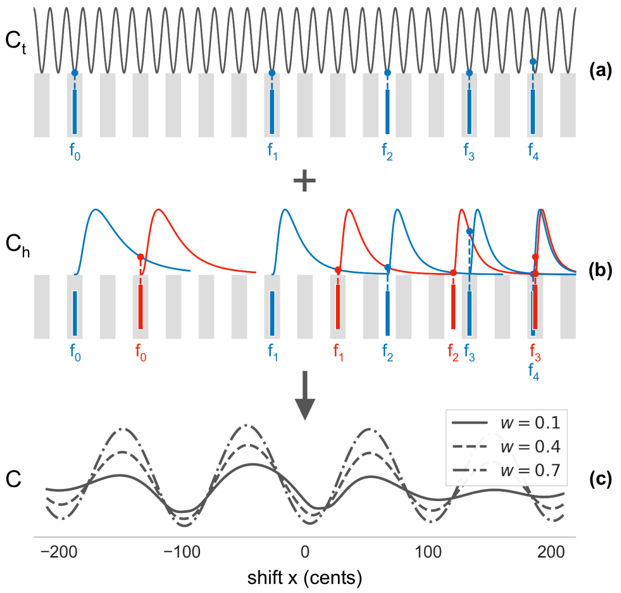

Figure 2: Conceptual overview of our cost measure. **(a)** C_t is higher when salient frequencies are far from the equal temperament grid. **(b)** C_h is higher when salient frequencies of a tone (blue) are similar but not equal to salient frequencies of a concurrent reference tone (red). **(c)** Shifting the tone shown in blue by x cents changes the overall cost C. A higher relative weight w of C_t to C_h results in local minima closer to the 12-TET grid.

A lower cost C corresponds to a "better" choice for the pitch, where the parameter $w \in [0, 1]$ controls the relative weight between the two aspects.

Figure 2 exemplifies the behavior of C for two hypothetical tones with five harmonic partials each (depicted in blue and red). The graphs in Figure 2c show the change in C when the tone represented in blue is pitch-shifted by -200 to $+200$ cents. As the two tones form a major third, a shift by $+14$ cents would result in a just interval. With a larger w, the relative weight of C_t increases, corresponding to a preference for equal temperament, whereas with decreasing w, the local minima move closer to JI intervals.

In the upcoming section, we develop differentiable expressions for C_t and C_h and illustrate their properties by continuing our example from Figure 1. In Section 3.4, we then show how the cost measure can be used to adapt the intonation using gradient descent.

3.1 Prerequisites

For a given audio signal, we assume a stationary sound in each analysis time frame n and represent it by a set

$$\mathcal{P}[n] := \{(f_m, a_m) \mid m \in \{1, ..., M\}\}, \tag{2}$$

consisting of M salient frequencies f_m in Hz with amplitude a_m. The cost measure is defined for a signal w.r.t. to a reference (or "background") signal represented by a set $\mathcal{P}_\mathrm{ref}[n]$. In the following, we omit the frame index for $\mathcal{P}$ and $\mathcal{P}_\mathrm{ref}$ where the time-dependency is not relevant.

To obtain this representation from audio signals, we use a short-time Fourier transform (STFT) with a frame and hop size of 0.1 sec and detect peaks in the magnitude spectrum of each time frame that constitute the salient frequencies. Avoiding the misinterpretation of transient peaks in the spectrum, we filter the STFT representation to remove percussive components of the signal [22]. Then, we identify up to 16 peaks in the remaining spectrum of each time frame by selecting the local maxima above a threshold. To increase frequency resolution, we interpolate the exact peak frequency and amplitude by fitting a parabola to the magnitudes of neighboring bands [23]. For an inactive voice or a purely percussive signal frame, we set $\mathcal{P} = \emptyset$.

Note that, for harmonic sounds in a monophonic signal, all salient frequencies in $\mathcal{P}$ are close to integer multiples of the lowest frequency f_0. This assumption does not hold for inharmonic sounds and polyphonic recordings.

For the example in Figure 1, we synthesize the signals using a sawtooth waveform with 16 harmonic partials and amplitudes $a_i = 1/(i\pi)$ using a reference frequency of 440 Hz for A4.

3.2 Tonal Cost

Equal divisions of the octave are a good approximation for the distribution of pitches in many music theories [11]. Furthermore, measuring the distance to an equal temperament grid is an often used strategy to assess the intonation in a performance [3, 24].

We define the tonal distance $d_\mathrm{t}^K(f_1, f_2)$ between two positive frequencies f_1 and f_2 in Hz on a K-TET grid as

$$d_\mathrm{t}^K(f_1, f_2) := \frac{1}{2}\Big(1 - \cos\big(2\pi K \log_2(f_1/f_2)\big)\Big), \tag{3}$$

where the distance is small if the interval between f_1 and f_2 is close to a K-TET interval (i. e., $f_1/f_2 \approx 2^{k/K}$ with $k \in \mathbb{Z}$). By measuring the tonal distance of each frequency in $\mathcal{P}$ to a given reference frequency f_ref, we define the tonal cost C_t as

$$C_\mathrm{t} := \frac{\sum_{(f,a)\in\mathcal{P}} a \cdot d_\mathrm{t}^K(f, f_\mathrm{ref})}{\sum_{(f,a)\in\mathcal{P}} a}, \tag{4}$$

where frequencies with higher amplitude contribute more to C_t. The highest cost $C_\mathrm{t} = 1$ is reached, when all salient frequencies lie exactly in the middle between two frequencies on the equal temperament grid defined by K and f_ref. The parameters K and f_ref can either be estimated from $\mathcal{P}_\mathrm{ref}$ or fixed to known values. Note that, with fixed parameters, C_t does not depend on $\mathcal{P}_\mathrm{ref}$. Furthermore, Figure 2a shows that for integer-multiple frequencies in $\mathcal{P}$, not all frequencies can align with the grid, so that C_t is never 0 for such signals. However, the local minimum is reached when the loudest partials are close to the grid.

For the example, let $\mathcal{P}$ include the salient frequencies of the soprano voice while all other voices are contained in $\mathcal{P}_\mathrm{ref}$. By setting $K = 12$ and $f_\mathrm{ref} = 440$ Hz, we obtain the cost heatmap shown in Figure 3. By pitch-shifting the soprano signal by -200 to 200 cents and evaluating the cost for each time frame, it shows for which shifts the salient frequencies in the signal fit best on the 12-TET grid. Tracking the nearest local minimum starting at a shift of 0

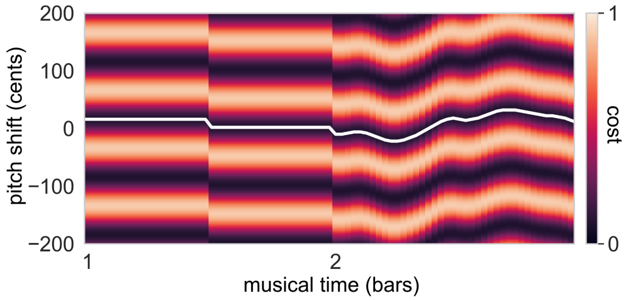

Figure 3: Tonal cost heatmap for the soprano voice in the example from Figure 1, pitch-shifted by -200 to $+200$ cents. White line indicates local minimum closest to 0.

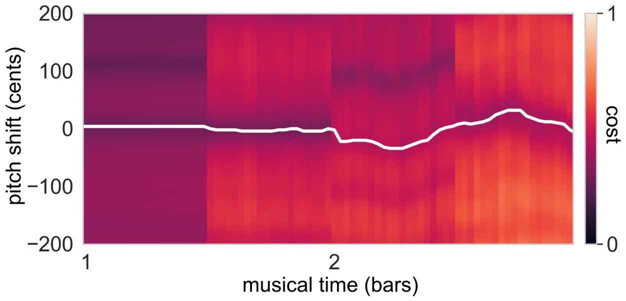

Figure 4: Harmonic cost heatmap for the soprano voice in the example from Figure 1, pitch-shifted by -200 to $+200$ cents. White line indicates local minimum closest to 0.

cents, the white line corresponds to the pitch shift required to minimize the cost.

3.3 Harmonic Cost

The perceptual dissonance between concurrent sounds can be expressed in terms of the *pure-tone* dissonance between all combinations of salient frequencies present in each sound [19]. While the perceived dissonance between two pure tones was first determined experimentally [20], we quantify the dissonance between two positive frequencies f_1 and f_2 using the parametrized model from [11] (omitting a global scaling factor):

$$d_{\mathrm{h}}(f_1, f_2) := \exp\Big(-\ln^2\big(\frac{|\log_2(f_1/f_2)|}{w_{\mathrm{c}}}\big)\Big), \tag{5}$$

with $w_{\mathrm{c}} := 6.7 \cdot \min(f_1, f_2)^{-0.68}$ as the frequency-dependent parameter that controls the interval of maximal dissonance and the decay of the dissonance curve. To ensure differentiability, we define $d_{\mathrm{h}}(f_1, f_2) := 0$ for $f_1 = f_2$.[2] As illustrated in Figure 5, $d_{\mathrm{h}}(f_1, f_2)$ approaches 0 from both sides when f_1 is close to f_2. $d_{\mathrm{h}}(f_1, f_2) = 1$ is maximal when the logarithmic distance between the two frequencies is $|\log_2(f_1/f_2)| = w_{\mathrm{c}}$.

The sum of $d_{\mathrm{h}}(f_1, f_2)$ between all pairs of salient frequencies in $\mathcal{P}$ and $\mathcal{P}_{\mathrm{ref}}$ weighted by the amplitude constitutes the harmonic cost:

$$C_{\mathrm{h}} := \frac{\sum\limits_{(f,a)\in\mathcal{P}} \; \sum\limits_{(f_{\mathrm{r}},a_{\mathrm{r}})\in\mathcal{P}_{\mathrm{ref}}} \min(a, a_{\mathrm{r}}) \cdot d_{\mathrm{h}}(f, f_{\mathrm{r}})}{\sum\limits_{(f,a)\in\mathcal{P}} a} \tag{6}$$

The normalization does not restrict C_{h} to $[0, 1]$, because the total number of salient frequency pairings is $|\mathcal{P}| \cdot |\mathcal{P}_{\mathrm{ref}}|$, but when comparing harmonic signals, only a small fraction of pairings have $d_{\mathrm{h}}(f_1, f_2) \gg 0$. By normalizing by the sum of amplitudes in $\mathcal{P}$, we achieve a comparable range for C_{t} and C_{h}, where C_{h} vanishes when the amplitudes in $\mathcal{P}_{\mathrm{ref}}$ are very small (i. e. no reference signal is present for which a harmonic relation can be evaluated).

The concept of the harmonic cost is illustrated in Figure 2b, where only the pairings between sets (blue and red) contribute to the cost. With the frequency-dependent w_{c}, the dissonance curve becomes more narrow towards higher frequencies. Applied to the soprano in our example analogous to Figure 3, this results in the heatmap in Figure 4.

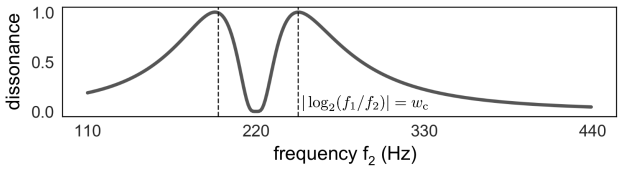

Figure 5: Pure-tone dissonance $d_{\mathrm{h}}(f_1, f_2)$ from [11] with $f_1 = 220$ Hz and f_2 between 110 and 440 Hz.

3.4 Joint Intonation Adaptation

In a musical performance, the cost C may vary between time frames n and we denote the cost in each frame by $C[n]$. The goal of intonation adaptation is to obtain a pitch shift function $p : \mathbb{Z} \to \mathbb{R}$, where a pitch shift of $p[n]$ cents applied to the considered signal minimizes $C[n]$.

When multiple voices can be adapted simultaneously in a polyphonic multitrack setting, the cost for each individual voice depends on the salient frequencies in the other voices. We denote the cost for each voice v with respect to all other voices by $C_v[n]$ and the current pitch shift for the signal of v by $p_v[n]$. Then the optimal shift can be found by solving

$$\min_{p_v[n]} \quad C_v[n]. \tag{7}$$

As described in [12], gradient descent is an effective method to find the local minimum. Pitch-shifting with $p_v[n]$ affects the salient frequencies in $\mathcal{P}[n]$ equally on a logarithmic scale while the frequencies in $\mathcal{P}_{\mathrm{ref}}[n]$ stay constant. Thus, the shift p in cents can be moved out of the logarithm in the tonal distance $d_{\mathrm{t}}^K(f_1, f_2)$ and the dissonance $d_{\mathrm{h}}(f_1, f_2)$, for which we introduce auxiliary functions $\delta_{\mathrm{t}}^K(f_1, f_2, p)$ and $\delta_{\mathrm{h}}(f_1, f_2, p)$:

$$\begin{aligned} \delta_{\mathrm{t}}^K(f_1, f_2, p) &:= \frac{1}{2}\Big(1 - \cos\big(2\pi K(\log_2(f_1/f_2) + p/1200)\big)\Big) \\ \delta_{\mathrm{h}}(f_1, f_2, p) &:= \exp\Big(-\ln^2\big(\frac{|\log_2(f_1/f_2) + p/1200|}{w_{\mathrm{c}}}\big)\Big) \end{aligned} \tag{8}$$

In the following, we omit the arguments of the distance and dissonance functions for brevity. Analogous to (5),

[2] Proof of differentiability can be found on the accompanying website.

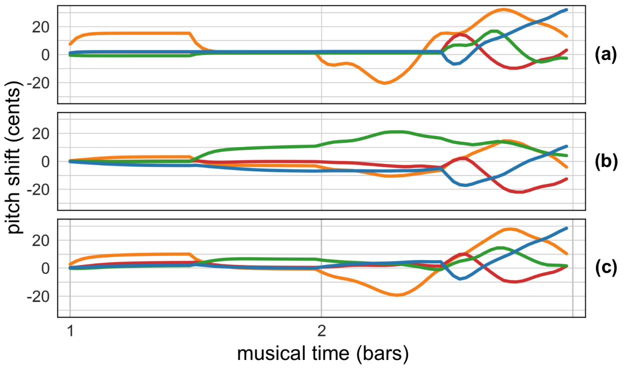

Figure 6: Adaptation curves $p_v[n]$ resulting from joint gradient descent with all four voices (orange: soprano, red: alto, green: tenor, blue: bass, $\mu = 350$). **(a)** $w = 1.0$ **(b)** $w = 0.0$ **(c)** $w = 0.33$

we define $\delta_{\mathrm{h}} := 0$ for $\log_2(f_1/f_2) = -p/1200$ to retain differentiability. Furthermore, we assume for δ_{h} that w_{c} stays constant for small shifts p, so that its frequency-dependency does not play a role in a single gradient descent step. Replacing d_{t}^K with δ_{t}^K and d_{h} with δ_{h} in (4) and (6) allows calculating the derivative of $C_v[n]$ directly with respect to p, so that the update rule becomes

$$p_{v,\text{new}}[n] = p_v[n] - \mu \frac{dC_v[n]}{dp}, \tag{9}$$

where μ is a step size parameter and the derivative of $C_v[n]$ (with the unit "cost change per cent shifted") is a weighted sum of $\frac{d\delta_{\mathrm{t}}^K}{dp}$ and $\frac{d\delta_{\mathrm{h}}}{dp}$:

$$\frac{d\delta_{\mathrm{t}}^K}{dp} = \frac{\pi K}{1200} \sin\left(2\pi K(\log_2(f_1/f_2) + p/1200)\right) \tag{10}$$

$$\frac{d\delta_{\mathrm{h}}}{dp} = -\frac{\ln(\frac{|\log_2(f_1/f_2)+p/1200|}{w_{\mathrm{c}}})}{600(\log_2(f_1/f_2) + p/1200)} \exp(-\ln^2(\frac{|\log_2(f_1/f_2) + p/1200|}{w_{\mathrm{c}}})). \tag{11}$$

with $\frac{d\delta_{\mathrm{h}}}{dp} = 0$ for $\log_2(\frac{f_1}{f_2}) = -\frac{p}{1200}$. By setting $p_v[n]$ to an initial value (e. g. 0) and repeatedly evaluating (9), we can now iteratively find the local minimum of $C_v[n]$ for each time frame. However, for short frame sizes, correlation between salient frequencies in successive time frames can be expected. Therefore, instead of finding the closest local minimum for each frame independently, we can use the pitch shift from the previous frame as the initial value for $p_v[n]$. Furthermore, to retain natural short-term pitch variations in the signal (e. g. vibrato), we require a certain temporal smoothness of $p_v[n]$. This can be achieved by updating $p_v[n]$ with only a single gradient descent step in each time frame, which yields

$$p_v[n] = p_v[n-1] - \mu \frac{dC_v[n]}{dp} \tag{12}$$

for $n > 0$ and $p_v[0] = 0$. Together with the frame size, the step size μ controls the rate at which pitch changes in the signals influence $p_v[n]$. This can be observed in Figure 6, which shows the resulting $p_v[n]$ from joint gradient descent with (12) for a varying weight w between C_{t} and C_{h} in the example cadence. For example, the pitch shift for the soprano voice visibly approaches a minimum in the first few frames of chords I and II. Moreover, it can be seen that the harmonic cost alone (Figure 6b) is not robust in musically dissonant chords like chord III, whereas with $w = 0.33$ (Figure 6c), the obtained pitch shift tends towards JI without ending up in a local minimum far away from equal temperament (cf. tenor in chords III and IV).

4. APPLICATION: INTONATION ADAPTATION IN CHORAL MUSIC

In the previous section, we introduced a method to obtain pitch shift curves by minimizing a cost measure that quantifies two aspects of intonation: the distance of a pitch to an equal temperament grid and the perceptual dissonance with regard to a harmonic reference. As a tool, this allows sound engineers to flexibly adapt the intonation in audio recordings between equal temperament and JI using the two parameters w and μ. With w, the relative weight between both aspects can be adjusted depending on artistic intent and musical context. μ controls the temporal behavior of the adaptation, where a larger μ corresponds to a stronger reaction to short-term pitch fluctuations.

A subjective evaluation of preferred intonation in different musical contexts and the resulting suitable choices for the parameters of our cost measure is beyond the scope of this paper. Many additional aspects, including timbre, acoustics, and performative choices (vibrato, portamento, etc.) [25], as well as listener taste and experience [26], influence intonation perception. Instead, we demonstrate the utility of the cost-based adaptation tool with an example from amateur performances of a cappella choral music. In this application with particularly volatile intonation, we show that the approach is robust on real-world signals and can blindly achieve results that are comparable to score-informed intonation adaptation.

For this, we apply the presented method to recordings from the *Dagstuhl ChoirSet* (DCS) [27]. The intonation adaptation of individual voices in a vocal recording requires separate signals for each voice and the dataset contains headset microphone signals for each singer. In this section, we consider the last four bars (45 to 48) from a performance of the motet *Locus Iste* (WAB 23, 1869) by Anton Bruckner (Quartet B, Take 3 in the dataset). The four-part a cappella composition is performed by a quartet of soprano (S), alto (A), tenor (T) and bass (B).

First, we obtain the salient frequencies for each voice from the four individual headset microphone signals, using the method described in Section 3.1. For the STFT, we keep the hop size of 0.1 sec (2205 samples in the DCS audio signals) and use a window size of 4096 samples for an improved frequency resolution. Due to varying levels and timbre of the singing voice and background noise, the number of salient frequencies in $\mathcal{P}[n]$ fluctuates. On average, $|\mathcal{P}[n]|$ is 6.6 (S: 5.6, A: 6.8, T: 4.4, B: 9.6) in the voiced frames (i.e., where $|\mathcal{P}[n]| > 0$). To assess the robustness of the detected salient frequencies against crosstalk and noise, we calculate the average deviation of each frequency from being an integer multiple of the lowest frequency in this frame (corresponding to the "harmonicity" of the signal). In the excerpt, the detected frequencies deviate from the harmonic overtones by a factor of 1.01 on average (S: 1.007, A: 1.012, T: 1.006, B: 1.016).

We now compute a pitch shift $p_v^{\mathrm{CB}}[n]$ for each voice with the cost-based (CB) method as described in Section 3.4, using a single gradient descent step for each frame (see (12)). For the present example, we set $w = 0.2$ and $\mu = 350$ and used fixed parameters $K = 12$ and $f_{\mathrm{ref}} = 440$ Hz for the tonal cost. The pitch shifts $p_v^{\mathrm{CB}}[n]$ are applied to each signal with a time-variant pitch shift algorithm based on resampling and time-scale modification [28]. In a cappella performances, one often observes (downward) intonation drifts [10], causing $p_v^{\mathrm{CB}}[n]$ to drift in the opposite direction to counteract this effect. For instance, the mean pitch shift across all voices in bar 48 of our example is 21 cents. Note that, to counteract a global drift of all voices in a similar direction, even $|p_v^{\mathrm{CB}}[n]| > 50$ cents may be intended. In this case, additional regularization can be added in the cost minimization to avoid individual voices ending up in local minima that do not reflect the relative intonation in the performance.

For the comparison of our method with a score-informed baseline (BL) approach, we estimate the F0 trajectories for each voice with pYIN [1] and assign the measurements to individual notes from the aligned score annotation provided in DCS. Then, for each time frame of 0.1 sec duration, we choose a pitch shift $p_v^{\mathrm{BL}}[n]$ that shifts the median F0 in the current frame onto the 12-TET pitch of the corresponding note in the score. To counteract larger fluctuations that result from the relatively small frame size for this method, we additionally smooth $p_v^{\mathrm{BL}}[n]$ using a moving average with a window size of 3 frames. The shift is applied to the signals in the same way as $p_v^{\mathrm{CB}}[n]$.

The pitch shift curves for the excerpt, calculated with the blind CB approach (colored) and the score-informed BL (black), are plotted for all voices in Figure 7a, c, e, and g. Furthermore, subfigures b, d, f, and h show the F0 trajectories of the original (black) and adapted (CB: colored, BL: grey) signals. The difference between the two pitch shift curves is small in most frames, particularly when compared to the magnitude of overall pitch fluctuations in the singing voices. Larger differences between the curves can be observed at the onset of some notes. This can be attributed to the strong influence of short-term fluctuations in the measured F0 trajectories on the BL approach.

In addition, the harmonic cost term C_{h} has a recognizable effect on the local minimum where JI intervals differ from 12-TET. This can be prominently observed in the soprano voice in bar 48 (c.f. the zoomed detail in Figure 7), where the sung note E is the major third of the final C major chord of the piece and therefore has a JI pitch 14 cents lower than 12-TET. This shows that our real-time capable method for cost-based intonation adaptation is able to approach JI tuning in vocal recordings without explicit knowledge about scales and keys.

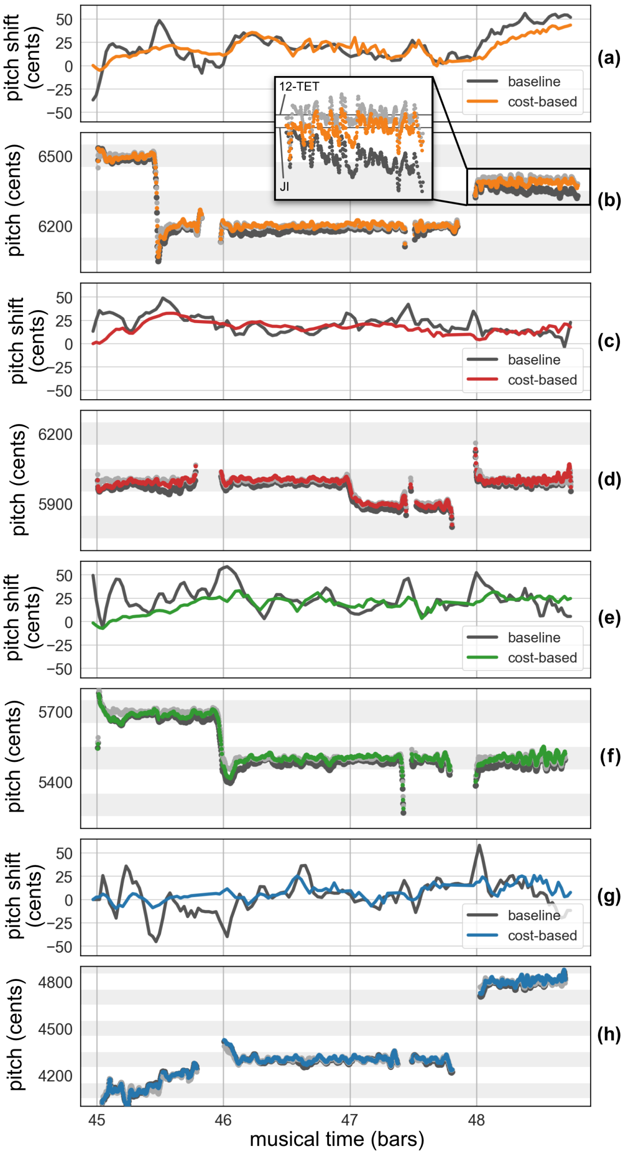

Figure 7: Joint adaptation of bars 45-48 of A. Bruckner "Locus Iste" (DCS, Quartet B, Take 3). The F0 trajectory plots (b,d,f,h) show the F0 of the original signal (black), the baseline (BL, grey) and the cost-based (CB, colored, $w = 0.2$, $\mu = 350$) pitch-shifted signals. **(a, b)** Soprano **(c, d)** Alto **(e, f)** Tenor **(g, h)** Bass

5. CONCLUSION

In this paper, we introduced a differentiable cost measure for intonation processing in polyphonic music recordings, which accounts for a tonal and a harmonic aspect in a user-specified proportion. Our method can be used as a flexible tool for intonation adaptation in multitrack choral music recordings. In future work, we will investigate the perceptual implications of our intonation adaptation in real-world signals. Furthermore, we want to apply this principle to more intonation processing tasks such as adaptive tuning of synthesizers and explore ways to incorporate additional aspects of intonation in the cost measure.

Acknowledgements: This project is supported by the German Research Foundation (DFG MU 2686/12-1, MU 2686/13-1). The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits IIS.

6. REFERENCES

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

[2] M. Mauch, K. Frieler, and S. Dixon, "Intonation in unaccompanied singing: Accuracy, drift, and a model of reference pitch memory," *Journal of the Acoustical Society of America*, vol. 136, no. 1, pp. 401–411, 2014.

[3] C. Weiß, S. J. Schlecht, S. Rosenzweig, and M. Müller, "Towards measuring intonation quality of choir recordings: A case study on Bruckner's Locus Iste," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Delft, The Netherlands, 2019, pp. 276–283.

[4] J. Devaney, M. I. Mandel, and I. 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)*, Porto, Portugal, 2012, pp. 511–516.

[5] F. Havrøy, "'You Cannot Just Say: "I am Singing the Right Note"'. Discussing intonation issues with Neue Vocalsolisten Stuttgart," *Music & Practice*, vol. 1, 2013.

[6] A. Grell, J. Sundberg, S. Ternström, M. Ptok, and E. Altenmüller, "Rapid pitch correction in choir singers," *The Journal of the Acoustical Society of America*, vol. 126, no. 1, pp. 407–413, 2009.

[7] J. Dai and S. Dixon, "Singing together: Pitch accuracy and interaction in unaccompanied unison and duet singing," *The Journal of the Acoustical Society of America*, vol. 145, no. 2, pp. 663–675, 2019.

[8] ——, "Intonation trajectories within tones in unaccompanied soprano, alto, tenor, bass quartet singing," *The Journal of the Acoustical Society of America*, vol. 146, no. 2, pp. 1005–1014, 2019.

[9] P.-G. Alldahl, *Choral Intonation.* Gehrmans Musikförlag, 1990.

[10] D. M. Howard, "Intonation drift in a capella soprano, alto, tenor, bass quartet singing with key modulation," *Journal of Voice*, vol. 21, no. 3, pp. 300–315, 2007.

[11] J. Berezovsky, "The structure of musical harmony as an ordered phase of sound: A statistical mechanics approach to music theory," *Science Advances*, vol. 5, p. eaav8490, May 2019.

[12] W. Sethares, "Adaptive tunings for musical scales," *The Journal of the Acoustical Society of America*, vol. 96, 1994.

[13] Celemony, "Melodyne," https://www.celemony.com/en/melodyne/, accessed: 2021-05-07.

[14] Antares, "Auto-Tune," https://www.antarestech.com/, accessed: 2021-05-07.

[15] D. Code, "Groven.Max: An adaptive tuning system for MIDI pianos," *Computer Music Journal*, vol. 26, pp. 50–61, 2002.

[16] W. Mohrlock, "Hermode Tuning," http://www.hermode.com/index_en.html, accessed: 2021-05-07.

[17] K. Stange, C. Wick, and H. Hinrichsen, "Playing music in just intonation: A dynamically adaptive tuning scheme," *Computer Music Journal*, vol. 42, no. 3, pp. 47–62, 2018.

[18] S. Wager, G. Tzanetakis, C. Wang, and M. Kim, "Deep autotuner: A pitch correcting network for singing performances," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).* Barcelona, Spain: IEEE, 2020, pp. 246–250.

[19] W. A. Sethares, "Local consonance and the relationship between timbre and scale," *Journal of the Acoustical Society of America*, vol. 94, no. 3, pp. 1218–1228, 1993.

[20] R. Plomp and W. J. M. Levelt, "Tonal consonance and critical bandwidth," *Journal of the Acoustical Society of America*, vol. 38, no. 4, pp. 548–560, 1965.

[21] J. Villegas and M. Cohen, "Roughness minimization through automatic intonation adjustments," *Journal of New Music Research*, vol. 39, pp. 75 – 92, 2010.

[22] D. FitzGerald, "Harmonic/percussive separation using median filtering," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, Graz, Austria, September 2010, pp. 246–253.

[23] J. O. Smith and X. Serra, "PARSHL: An analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation," in *Proceedings of the International Computer Music Conference (ICMC).* Computer Music Association, 1987.

[24] T. Nakano, M. Goto, and Y. 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)*, Pittsburgh, PA, USA, 2006, pp. 1706–1709.

[25] J. M. Geringer, R. B. MacLeod, C. K. Madsen, and J. Napoles, “Perception of melodic intonation in performances with and without vibrato,” *Psychology of Music*, vol. 43, no. 5, pp. 675–685, 2015.

[26] F. Loosen, “The effect of musical experience on the conception of accurate tuning,” *Music Perception*, vol. 12, no. 3, pp. 291–306, 1995.

[27] S. Rosenzweig, H. Cuesta, C. Weiß, F. Scherbaum, E. Gómez, and M. Müller, “Dagstuhl ChoirSet: A multitrack dataset for MIR research on choral singing,” *Transactions of the International Society for Music Information Retrieval (TISMIR)*, vol. 3, no. 1, pp. 98–110, 2020.

[28] S. Rosenzweig, S. Schwär, J. Driedger, and M. Müller, “Adaptive pitch-shifting with applications to intonation adjustment in a cappella recordings,” to appear in Proceedings of the International Conference on Digital Audio Effects (DAFx), Vienna, Austria, 2021.