Multi-Voice Intonation Adaptation via Gradient Descent

Simon Schwär, *Graduate Student Member, IEEE*, and Meinard Müller, *Fellow, IEEE*

***Abstract*—Intonation in ensemble performances on instruments with flexible tuning involves a complex interaction between musicians shaped by musical context, acoustic conditions, and each performer's perception and preferences. In post-production, it is often desirable to compensate for unintended deviations while preserving expressive fluctuations and musically meaningful interaction between individual tracks and voices. In this paper, we formulate multi-voice intonation adaptation as a cost minimization problem, making three main contributions. First, we introduce a differentiable cost function that explicitly balances the adherence of each voice to an equal-tempered pitch grid and the harmonic fit between voices using sensory dissonance. Second, based on this differentiable cost function, we derive a gradient descent adaptation algorithm that produces smooth, time-varying pitch-shift curves without requiring score or note-level information. We show how a small set of interpretable hyperparameters, including initialization, stopping criterion, step size, and momentum, allows for a controlled trade-off between the compensation of unintended intonation deviations and the preservation of expressive fluctuations. Third, we evaluate our method on string, wind, and vocal quartet multi-track recordings through objective and subjective experiments, demonstrating quality comparable to a commercial pitch-correction baseline while offering particular advantages in handling intonation drift and modeling interactions between voices. Beyond these results, the focus of this work is conceptual, making a typically heuristic post-production task transparent and controllable through an explicit cost-based optimization framework.**

***Index Terms*—Intonation, pitch adaptation, sensory dissonance, cost measure, gradient descent.**

I. Introduction

THE widely-used 12-tone equal temperament (12-TET) tuning system divides the octave into twelve semitones that are equally spaced on a logarithmic frequency axis. This allows musical instruments with fixed pitch, such as the piano, to play in any key. While 12-TET has become the dominant tuning system in Western music since the 18th century [1], the irrational ratio of $2^{1/12} \approx 1.0595$ between semitones implies that most intervals deviate slightly from the natural overtone relationships of harmonic sounds. In contrast, just intonation (JI) scales are explicitly constructed from intervals with small integer ratios, yielding tunings that align more closely with harmonic spectra. However, the absolute pitch values in JI scales depend on the chosen root note, requiring adaptation across keys and musical contexts.

Many musical instruments allow for continuous pitch variation or at least offer some flexibility in intonation[1]. This enables musicians to dynamically adjust their pitch to achieve

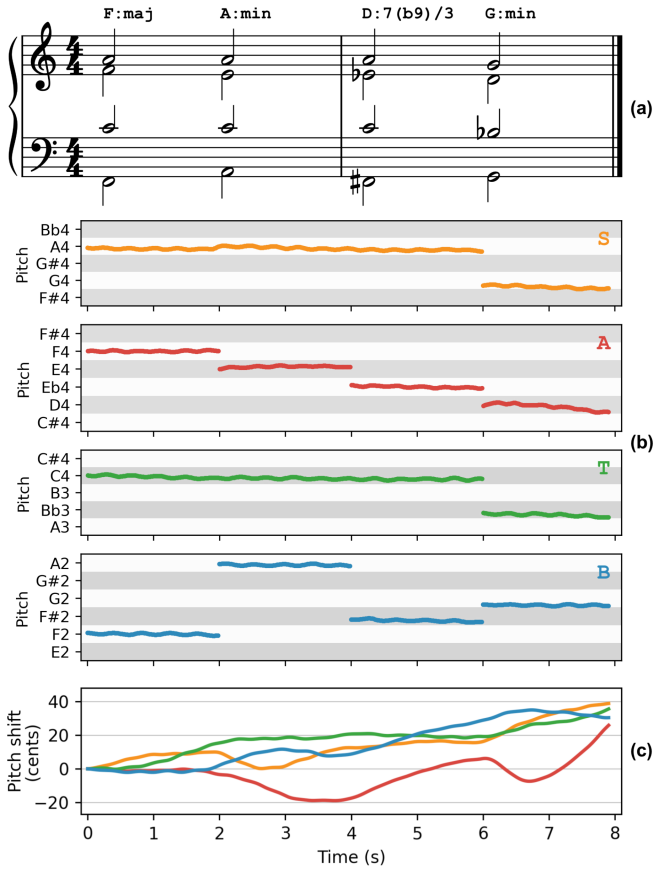

Fig. 1. Synthesized example cadence to illustrate aspects of intonation. (a) Sheet music with chord labels. (b) Fundamental frequencies (F0) of soprano (S, orange), alto (A, red), tenor (T, green), and bass (B, blue). (c) Pitch shift curve $p_v(n)$ for all voices $v \in \{\mathtt{S}, \mathtt{A}, \mathtt{T}, \mathtt{B}\}$ obtained with multi-voice gradient descent (hyperparameters $w = 0.33$, $L = 1$, $\mu = 50$, $\beta = 0.9$).

Received 23 July 2025; revised 18 December 2025 and 10 March 2026; accepted 14 March 2026. Date of publication 19 March 2026; date of current version 7 May 2026. This work was supported by German Research Foundation (DFG MU 2686/13-2) under Grant 401198673. The associate editor coordinating the review of this article and approving it for publication was Prof. Juhan Nam. *(Corresponding author: Simon Schwär.)*

The authors are with the International Audio Laboratories (a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits IIS), 91058 Erlangen, Germany (e-mail: simon.schwaer@audiolabs-erlangen.de; meinard.mueller@audiolabs-erlangen.de).

Digital Object Identifier 10.1109/TASLPRO.2026.3675812

[1]Here and in the following, we refer to a set of pitches that defines a musical scale as *tuning* and the realization of tuning by a performer as *intonation*.

a context-dependent intonation beyond equal-tempered scales. In choir singing [2] or string ensembles [3], for example, an adaptation in response to concurrently sounding voices is often referred to as *vertical intonation* and JI intervals with minimal beating between overtones are considered desirable [4]. However, performers constantly have to balance this aspect with other intonation constraints like melodic intervals (*horizontal intonation*) and expressive pitch deviations such as vibrato [5], [6], [7] in their own and other voices. Together with practical limitations in pitch control and perception, this interaction can lead to subtle but perceptible unintended intonation deviations, including intonation *drift*, i.e., the gradual, systematic deviation from an original reference pitch.

Fig. 1 illustrates these phenomena in a synthesized four-part cadence. The score in Fig. 1(a) specifies soprano (S), alto (A), tenor (T), and bass (B) voices with chord labels following [8]. The rendered performance in Fig. 1(b) reveals several characteristic effects. All voices exhibit a global downward drift of approximately 50 cents relative to the 12-TET grid while largely preserving their relative intervals. In the S voice, the repeated A4 notes are each played with different intonation, demonstrating a vertical intonation depending on the harmonic context. In addition, all voices show local expressive variations such as vibrato. Despite all these deviations, the harmonic structure remains perceptually stable.[2]

In the post-processing of multi-track recordings, a sound engineer may attempt to correct unintended intonation deviations via pitch shifting. However, this requires defining a target pitch that, analogous to live performance, depends on multiple interacting criteria. For example, one may wish to compensate for global intonation drift, often regarded as a flaw in a cappella singing [9], [10], while preserving expressive fluctuations and context-dependent vertical alignment. Fig. 1(c) illustrates pitch-shift curves that achieve such drift compensation while retaining vibrato and chord-dependent pitch deviations that improve harmonic fit. Designing algorithms that automatically estimate time-varying pitch-shift trajectories without manual intervention, while still offering musically meaningful control, remains a challenging task.

In this paper, we build upon our previous work [11] and formulate intonation adaptation with multiple voices as a cost minimization problem. This can be seen as a generalization of commonly used grid-based methods, for which the target pitches have to be predefined, whereas a cost-based formulation allows to flexibly and dynamically account for different influences on intonation, such as the interaction between voices. In this context, we make three main contributions. First, we propose a differentiable cost measure that combines two terms which are designed to have local minima at pitch shifts where a voice may reach desirable intonation. The first term captures a *tonal* aspect by encouraging proximity to an equal-tempered pitch grid, reflecting the widespread use of equal temperament as an approximation of tonal organization in Western music [12]. The second term models a *harmonic* aspect by measuring the sensory dissonance [13], [14] between simultaneously sounding voices, thereby encouraging JI-like vertical alignment. With a weighted combination of these terms, our method enables flexible intonation adaptation between equal-tempered anchoring and harmonic adaptation. The tonal term ensures stable local minima even in harmonically complex or unstable chords, while the harmonic term allows musically meaningful deviations from the grid when supported by the ensemble context.

As a second main contribution and extension of [11], we introduce a gradient-based adaptation algorithm that minimizes the intonation cost to compute smooth, time-varying pitch-shift curves for each voice. By utilizing established methods from gradient-based optimization in this novel context, we show how hyperparameters such as the initial value, step size, stopping criterion and momentum provide direct control over the adaptation, ranging from a global drift compensation that preserves vibrato to more fine-grained local adaptation. The differentiability of the proposed cost function is central to this framework. Beyond gradient-based optimization, it is compatible with automatic differentiation tools and gradient-based learning frameworks. This opens perspectives for integrating perceptually motivated intonation objectives into hybrid signal-processing and deep learning systems.

As the third main contribution, we evaluate our approach using case studies on diverse multi-track recordings of string, wind, and a cappella vocal quartets. In comparison to existing commercial methods for intonation adaptation, our approach shows advantages in counteracting intonation drift and enables a flexible weighting of intonation aspects depending on user preference. In addition to our methodological contributions, a small-scale listening test suggests that combining both cost terms can yield improved results compared to using each term individually, with performance comparable to a commercial baseline. Given the inherently subtle and often subjective nature of intonation, we do not claim that our method is perceptually superior to these baselines. Rather, the primary strength of the approach lies in the flexible optimization framework that replaces heuristic post-production workflows with an explicit cost-based formulation.

The remainder of this article is structured as follows. Section II reviews related work. Section III introduces the differentiable intonation cost measure. Section IV presents the gradient-based multi-voice adaptation algorithm and analyzes its parameters. Section V reports experiments and listening results. Audio examples and a Python toolbox for intonation processing are provided online.[2]

II. Related Work

The intonation adaptation strategies implemented in many commercial products like *Melodyne* [15], *AutoTune* [16], or *Flex Pitch* [17] are typically based on an estimate of the F0 trajectory in a monophonic recording, with Melodyne also supporting F0 estimation for polyphonic instruments such as the piano. After (manually or semi-automatically) segmenting the F0 trajectory into individual notes, the signal is pitch-shifted so that the average F0 for each note approaches a predefined target

[2]Audio examples and code are available online: https://audiolabs-erlangen.de/resources/MIR/2026-TASLP-Intonation

value. This target can be chosen manually by the user or determined automatically from a predefined scale or musical score. Additional heuristics are often used to preserve expressive pitch variations such as portamenti or vibrato, while accounting for the dynamic interaction between voices would typically require manual intervention. However, these approaches generally do not account for the interaction between multiple voices, which would require manual intervention with these tools.

Dynamically tuning individual tones of simultaneously played intervals and chords has been explored in the context of digital synthesizers. Rule-based algorithms like *Groven.Max* [18] or *Hermode Tuning* [19] analyze the current chord and adjust the F0 for each voice to approximate JI, compromising when just intervals between all pairs of notes are impossible. This problem can also be addressed by solving a quadratic program [20], distributing the deviation from JI evenly across the voices, with auxiliary constraints for temporal continuity within voices. These approaches operate primarily on symbolic representations or synthesized tones and assume explicit chord knowledge, making them less directly applicable to multi-track audio recordings with existing intonation deviations.

Data-driven approaches aim to infer favorable tuning from training examples and have mainly been explored for singing voice adaptation. A desired F0 trajectory can be directly copied from a time-aligned reference recording [21], [22], or predicted from a score input using neural networks [23], [24]. A key challenge is the limited availability of suitable training data, ideally consisting of matched recordings with varying intonation quality. In practice, heuristic strategies are often used for dataset construction [25] or for designing training objectives, for example by learning to synthesize natural pitch variations while minimizing deviations from a target pitch input [23]. Moreover, few data-driven methods explicitly consider harmonic context. Wager et al. [26] condition pitch correction on a backing track rather than a fixed score. Hyodo et al. [27] propose a polyphonic vocal synthesizer trained on the relative intervals between voices rather than absolute F0 values. However, these approaches implicitly encode adaptation behavior in model weights, offering limited controllability to accommodate individual artistic intent.

Sethares [28] established a relationship between tuning and timbre of simultaneous sounds using the concept of sensory dissonance [13], [14], which is largely influenced by the perception of *roughness* due to beating between tonal components (e.g., partials). By evaluating pairwise interactions between partials, a *dissonance landscape* can be obtained, in which, if the tones have only harmonic partials, local minima occur at intervals with small integer ratios. This principle has been explored for adaptive tuning using gradient descent [29], where the F0 of tones is updated to minimize sensory dissonance, achieving a tuning similar to JI without requiring explicit musical analysis of the chords. However, depending on the number and amplitudes of the partials, not all intervals that are commonly used in Western music have clearly defined local minima in the dissonance landscape (see, e.g., MIDI pitches 59–61 in Fig. 2(c)), limiting robustness for more musically complex chords. To mitigate this, Villegas and Cohen [30] introduce a proximity constraint, allowing the F0 to only deviate by a few cents from the initial tuning. While this stabilizes the optimization locally, it does not allow for effectively addressing larger deviations such as global intonation drift. To our knowledge, no previous work formulates multi-voice intonation adaptation for recorded ensemble signals as a fully differentiable optimization problem that explicitly trades off adherence of each voice to an equal-tempered pitch grid and the harmonic fit between multiple voices combined with controllable temporal behavior through interpretable optimization parameters. The approach proposed in the following addresses this gap.

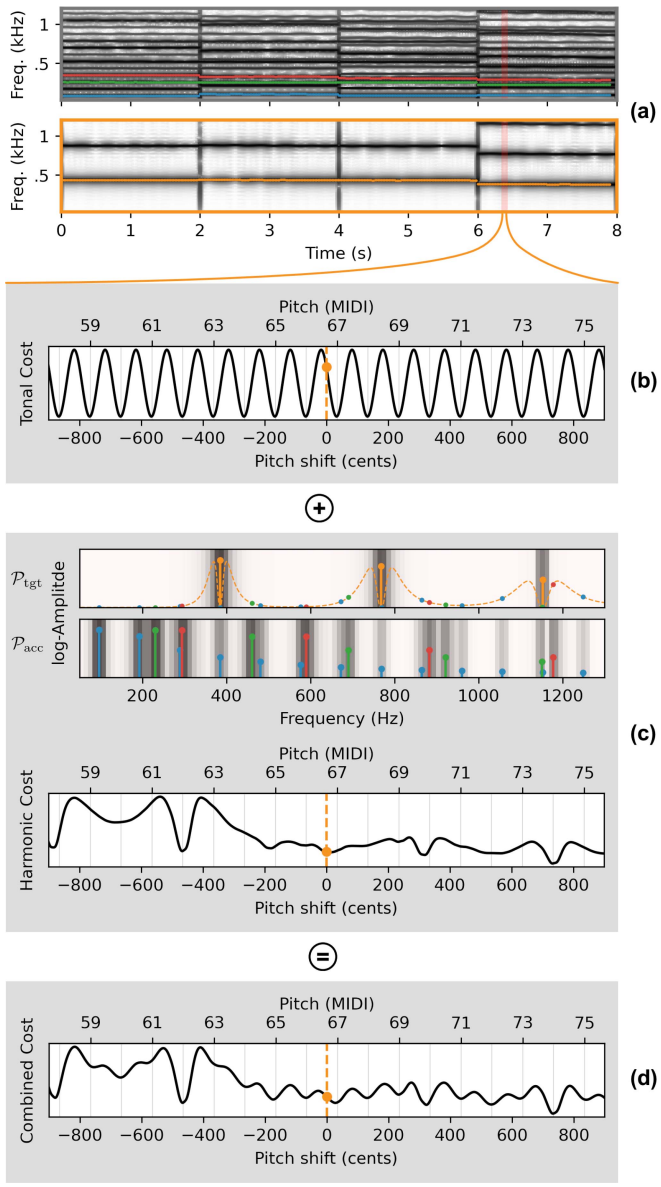

Fig. 2. Intonation cost terms for a single frame in the synthesized running example from Fig. 1. (a) Spectrogram of the accompaniment (ATB) and target voice (S) input signals with estimated F0 trajectories. (b) Tonal cost C_{t} using f_0^{tgt}. (c) $\mathcal{P}^{\mathrm{tgt}}$, $\mathcal{P}^{\mathrm{acc}}$ and harmonic cost C_{h} for the S voice in the highlighted frame. (d) Combined cost C with $w = 0.33$.

III. Intonation Cost Measures

By formulating intonation assessment as a cost measure, we obtain a flexible way to combine and balance multiple intonation criteria within a single mathematical framework. In particular, we define an *intonation cost* C for a monophonic *target voice* in the context of a fixed, potentially polyphonic *accompaniment*. For concise notation, we first consider a simplified setting with a single target voice and a single time frame. This is then extended in Section IV, where we will take into account the temporal relations over consecutive time frames as well as the interaction between multiple voices that can be adapted simultaneously.

To model how the target voice adapts and how this affects the intonation cost, we introduce a parameter p representing a pitch shift of the target voice in cents. Since we later want to compute the gradient $\partial C/\partial p$, we treat p as a direct parameter of the cost function, yielding a function $C : \mathbb{R} \to \mathbb{R}_+$, where $C(p)$ quantifies the intonation cost when the target voice is pitch-shifted by p cents. In the context of our adaptation algorithm, a suitable cost function C should satisfy two requirements: it should be differentiable with respect to p to enable gradient-based optimization, and it should attain local minima at pitch shifts p that are favorable under the chosen intonation criteria. In this work, we consider two such criteria, each represented by a separate term in the overall intonation cost. The *tonal* cost C_t measures the distance to an equal temperament (ET) grid, while the *harmonic* cost C_h measures sensory dissonance in the overall sound. To account for different artistic intents, both terms can be balanced via a weighting factor $w \in [0, 1]$, favoring dissonance minimization with $w \approx 0$ and adherence to ET with $w \approx 1$. This results in a combined cost function

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

The individual terms C_t and C_h reflect two key criteria of intonation [31], but obviously do not exhaustively capture all aspects that influence the perception of "good" intonation. For example, the combined cost does not consider horizontal intonation, which could be included by adding a third cost term that analyzes the intervals between consecutive time frames within the target voice. Our cost-based framework and adaptation method (see Section IV) allows for such extensions, as long as they preserve differentiability and yield local minima at favorable pitch shifts.

A. Signal Representations

In addition to the pitch shift p, the cost terms depend on the *tonal components* (or *partials*) of the target and/or accompaniment signals. For a given time frame, we represent an audio signal by the set

$$\mathcal{P} := \{(f_0, a_0), \ldots, (f_{M-1}, a_{M-1})\} \tag{2}$$

of M tonal components with frequency $f_m \in \mathbb{R}_+$ in Hz and amplitude $a_m \in \mathbb{R}_+$ for each component $m = 0, \ldots, M - 1$.

In general, $\mathcal{P}$ can be obtained from an audio signal by detecting prominent peaks in the signal's magnitude spectrum, where the components are not necessarily related by integer multiples. For monophonic and (approximately) harmonic sounds, such as the individual monophonic tracks in our running example, f_0 corresponds to the F0 of the tone, and $f_m \approx m f_0$ represent the integer-multiple partials. In this case, a more robust alternative to general peak finding is to first estimate the F0 and then identify spectral peaks at (or close to) its integer multiples [32], which reduces sensitivity to background noise and other spurious spectral peaks.

For the following definitions of the cost terms, we denote the salient tonal components of the monophonic target signal by $\mathcal{P}^\mathrm{tgt}$ and those of the accompaniment by $\mathcal{P}^\mathrm{acc}$. In our case, $\mathcal{P}^\mathrm{acc}$ is the union of the peaks identified from the individual accompaniment voices. We further denote the F0 from $\mathcal{P}^\mathrm{tgt}$ by f_0^tgt to simplify notation. An example of $\mathcal{P}^\mathrm{tgt}$ and $\mathcal{P}^\mathrm{acc}$ is illustrated in Fig. 2(a), using S from the example cadence (cf. Fig. 1) as the target voice and the mixture of all other voices (ATB) as the accompaniment.

B. Tonal Cost

The tonal cost C_t aims to quantify how far the target voice, represented by f_0^tgt, deviates from an equal-tempered (ET) pitch grid. In general, we denote a grid with K divisions of the octave as K-TET. In Western music, $K = 12$ (the 12-TET grid with twelve pitches per octave) is widely used to assess intonation quality, both in literature [33], [34], [35] and in practice (e.g., using chromatic tuners). While ET is sometimes seen as a compromise for instruments with fixed tuning [1], it provides a reasonable approximation of the tonal organization principle underlying many musical traditions [12].

To measure the deviation of f_0^tgt from an ET grid, we first introduce a function that takes two frequencies f and f' and measures how closely the interval between them matches an ET interval. Since the ET grid is evenly spaced on a logarithmic frequency axis, a suitable distance function should have periodic local minima and increase smoothly as the interval deviates from the grid. A cosine over the log-interval satisfies this requirement and is continuously differentiable, making it a suitable choice for modeling the tonal cost. We thus define the distance of the interval between f and f' to a K-TET interval by

$$d_\mathrm{t}^K(f, f') := \frac{1}{2}\left(1 - \cos\left(2\pi K \log_2(f/f')\right)\right). \tag{3}$$

This distance d_t^K is zero if the interval between f and f' is exactly a K-TET interval, i.e., $f/f' = 2^{k/K}$ with $k \in \mathbb{Z}$. With this, we define C_t as

$$C_\mathrm{t}(p) := d_\mathrm{t}^K(f_0^\mathrm{tgt} \cdot 2^{p/1200}, f_\mathrm{ref}), \tag{4}$$

where p is the pitch shift parameter defined above. The hyperparameters K and f_ref define the ET grid and can either be estimated from the accompaniment (see, e.g., [33]) or set to predefined values. In our experiments, we set $K = 12$ and $f_\mathrm{ref} = 440$ Hz (corresponding to standard A4 tuning).

The resulting tonal cost C_t is illustrated in Fig. 2(b). In the depicted frame, a pitch shift of $p = 0$ for the S voice leads to a non-zero cost, meaning that the original intonation does not lie exactly on the 12-TET grid. The nearest local minimum is exactly at the center frequency of MIDI pitch 67 which matches the target pitch (G4) defined by the score shown in Fig. 1(a). Since C_t has a minimum at each integer MIDI pitch, the nearest local minimum would be different from the score target if the

voice deviates by more than 50 cents. Note that C_t depends only on f_0^tgt and therefore neither considers differences in timbre (in particular, the energy distribution across partials) nor any interaction between target and accompaniment.

C. Harmonic Cost

The harmonic cost C_h aims to quantify the dissonance between simultaneously sounding tones. Following [13], we consider sensory dissonance as a differentiable surrogate for this perceptual property. In contrast to C_t, which evaluates the target voice in isolation against a fixed pitch grid, C_h explicitly models the interaction between the target voice and the accompaniment. It is important to note that sensory dissonance is not inherently a negative property, since its variation across different chords and timbres is an important component of musical expression [36]. However, when tones have only harmonic partials, locally minimizing sensory dissonance can lead to intonation similar to JI [29], which may be desirable in some ensemble performance. For chords consisting of more complex intervals, as well as inharmonic tones (such as many metallophones), the minima can differ considerably from typical intervals in Western harmony [13], which is a key limitation of this approach compared to the tonal stability offered by C_t.

The sensory dissonance between two complex simultaneous sounds can be expressed as the sum of the *pure-tone* dissonance between all combinations of tonal components in each sound [28]. This pure-tone dissonance was initially determined from perceptual experiments with simultaneously played sinusoids [37] and has since been parameterized in a multitude of different models (see [14] for an overview). We adopt the model from [12] for two reasons. First, it yields a smooth, continuously differentiable function (cf. Appendix A), which is a prerequisite for our cost terms as outlined above. Second, it is based on a logarithmic ratio $\log_2(f/f')$ between frequencies f and f' instead of a linear distance as many other models, simplifying the computation of the gradient $\partial C_\mathrm{h}/\partial p$. The pure-tone dissonance between two sinusoids is given by

$$d_\mathrm{h}(f,f') := \begin{cases} \exp\left(-\ln^2\left(\frac{\left|\log_2(f/f')\right|}{d_\mathrm{max}(\bar{f})}\right)\right) & f \neq f' \\ 0 & f = f', \end{cases} \tag{5}$$

where $d_\mathrm{max}(\bar{f})$ is a frequency-dependent parameter controlling the interval of maximal dissonance and the decay of the dissonance curve, which in [12] is defined based on a heuristic that increases d_max towards higher frequencies. To make this parameter more interpretable, we instead express $d_\mathrm{max}(\bar{f})$ explicitly in terms of the *equivalent rectangular bandwidth* (ERB) [38], a standard measure for auditory frequency resolution [39], placing the dissonance maximum precisely at 0.25 critical bands as established by perceptual studies [37], [40]:

$$d_\mathrm{max}(\bar{f}) := \log_2\left(1 + 24.7\left(0.00437\bar{f} + 1\right)/\bar{f}\right). \tag{6}$$

Here, the ERB is evaluated in octaves at the mean frequency $\bar{f} := (f + f')/2$, ensuring symmetry (i.e., $d_\mathrm{h}(f,f') = d_\mathrm{h}(f',f)$). As illustrated in Fig. 3, d_h approaches zero when the two frequencies are very similar or very far apart, and reaches its maximum

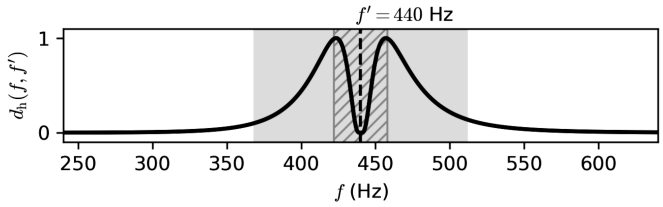

Fig. 3. Pure-tone dissonance $d_\mathrm{h}(f,f')$ for $f = 440$ Hz. The shaded area indicates the ERB around f (0.25 ERB marked with hatched rectangle).

of one at $|\log_2(f/f')| = d_\mathrm{max}(\bar{f})$. In the derivative of d_h, we assume $d_\mathrm{max}(\bar{f})$ to be constant, since it varies only minimally for small changes in $\bar{f}$.

The harmonic cost is then defined as the amplitude-weighted sum of pure-tone dissonances between all pairs of tonal components in $\mathcal{P}^\mathrm{tgt}$ and $\mathcal{P}^\mathrm{acc}$:

$$C_\mathrm{h}(p) := \frac{1}{A} \sum_{\substack{(f,a)\in\mathcal{P}^\mathrm{tgt} \\ (f',a')\in\mathcal{P}^\mathrm{acc}}} \min(a,a') d_\mathrm{h}(f \cdot 2^{p/1200}, f'), \tag{7}$$

where $\min(a,a')$ weights each pairing proportionally to the amplitude of the beating that can occur between the two components [13]. The normalization factor A ensures that C_h and C_t are in a comparable range. Normalizing by the sum of all weighting factors would guarantee $C_\mathrm{h} \in [0,1]$, but since most pairs of tonal components are further apart than one critical band, their contribution to the sum is negligible ($d_\mathrm{h} \approx 0$), resulting in a very small expected value for C_h. We therefore normalize only by the sum of weighting factors for pairs where $|\log_2(f/f')| \leq 1$, i.e., within one octave, which yields magnitudes comparable to those of C_t.

In Fig. 2(c), the harmonic cost C_h is illustrated for a single frame from our example cadence. The upper plot shows $\mathcal{P}^\mathrm{tgt}$ and $\mathcal{P}^\mathrm{acc}$ in this frame, along with the pure-tone dissonance between each pairing. As an example, consider the third harmonic of the S voice at around 1200 Hz: d_h is non-negligible for its pairing with the fourth harmonic of the A voice, as well as the eleventh and thirteenth harmonics of the B voice. Summing all these interactions yields the harmonic cost shown in the lower plot for pitch shifts p between -1100 and 800 cents. In contrast to C_t in Fig. 2(b), C_h exhibits a local minimum very close to the actual F0 of the target voice, reflecting the advantage of modeling voice interactions explicitly. On the other hand, C_h does not necessarily have a clear local minimum at every musically relevant target pitch. For example, near MIDI pitches 59 and 61 in Fig. 2(c), the nearest local minimum is close to MIDI pitch 60, illustrating the complementary role of C_t in ensuring tonal stability.

Combining both cost terms as in Fig. 2(d) allows for a flexible weighting of the two modeled intonation aspects. With $w = 0.33$ in this example, there is a clear minimum near most MIDI pitches, and this minimum is shifted towards reduced sensory dissonance where C_h exhibits a clear local minimum on its own.

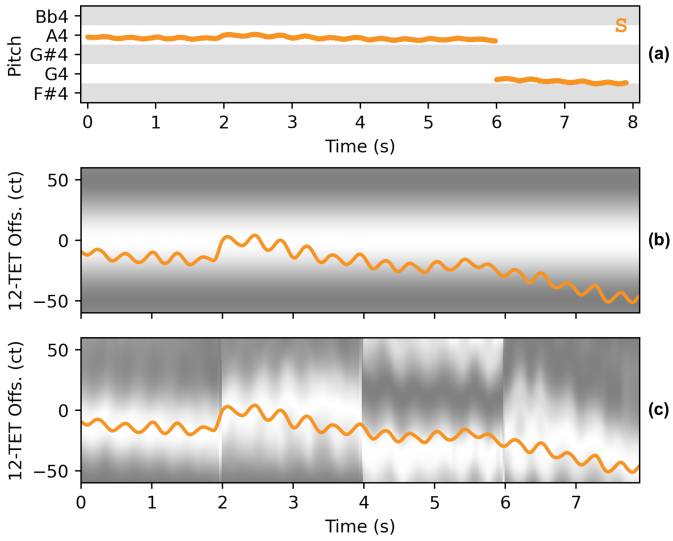

Fig. 4. Illustration of the intonation cost over time for the S voice of the example cadence from Fig. 1. (a) F0 of the S voice, reproduced from Fig. 1(b) for reference. (b) Visualization of cost term C_{t}. The y-axis represents the 12-TET offset, i.e., how far the F0 of the S voice deviates from the nominal 12-TET pitch of the played note. The actual deviation of the S voice is shown with an orange line. A darker background color denotes a higher cost. (c) Like (b), but visualizing cost term C_{h}.

IV. Intonation Adaptation Algorithm

The differentiability of C with respect to p enables a gradient-based adaptation algorithm that computes a pitch shift curve $p(n)$, specifying for each time frame n by how many cents the target voice should be pitch-shifted to reduce the overall intonation cost. By controlling the amount of adaptation applied in each frame, the algorithm can be tuned to correct global intonation deviations while preserving local ones, or vice versa, resulting in a more natural-sounding adaptation than a purely frame-wise correction would yield. This behavior can be controlled by several hyperparameters described below, which can be adjusted to match the properties of the input signal (e.g., the strength of intended and unintended pitch variations) as well as the desired output (e.g., a natural-sounding or strongly modified pitch trajectory). The resulting time-varying pitch shift $p(n)$ can then be applied to the target signal using a suitable pitch shifting algorithm (see, e.g., [41]). Extending the cost function to time-varying signals, we further account for the frame-wise changing properties of the input signals, denoted by $f_0^{\mathrm{tgt}}(n)$, $\mathcal{P}^{\mathrm{tgt}}(n)$, and $\mathcal{P}^{\mathrm{acc}}(n)$, and write the corresponding intonation cost as $C(p, n)$.

To analyze the behavior of the cost terms and the adaptation algorithm over time, we express the pitch shift p relative to the nominal 12-TET pitch of the currently played note, referring to this quantity as the *12-TET offset*. Fig. 4 shows this offset for the S voice of our example cadence, overlaid on the *cost landscape*, which depicts the intonation cost as a function of the 12-TET offset over time, with darker colors indicating higher cost. As expected, the landscape of the tonal cost C_{t} (Fig. 4(b)) is constant over time and identical across tones, since it depends only on the deviation from the ET grid and not on the musical context. The landscape of the harmonic cost C_{h} (Fig. 4(c)), on the other hand, is strongly influenced by the musical context, as reflected in three observable phenomena. First, the chord function of the target voice affects the position of the local minimum. Since S is playing the major third in the first chord, its local minimum lies lower than in the second chord, consistent with JI tuning where the major third is 14 cents below 12-TET, whereas the same pitch A4 serves as the root in the second chord and therefore has a local minimum around the 12-TET offset of 0 cents. Second, a gradual downward drift of the local minima across later notes reflects a global intonation drift present throughout the entire example. Third, the vibrato of the accompanying voices manifests as small periodic fluctuations in the position of the local minimum.

A. Gradient Descent

In [29], gradient descent is used to update the frequency of an individual salient tonal component directly. Inspired by this approach, we formulate an iterative update rule for the pitch shift p with L total iterations ($\ell = 1, \ldots, L$)

$$p^{(\ell)}(n) = p^{(\ell-1)}(n) - \mu g^{(\ell)}(n), \tag{8}$$

where

$$g^{(\ell)}(n) = \left.\frac{\partial C(p, n)}{\partial p}\right|_{p = p^{(\ell-1)}(n)} \tag{9}$$

is the gradient of the cost function $C(p, n)$ evaluated at $p^{(\ell-1)}(n)$ and μ is a step size parameter. The unit of $g^{(\ell)}(n)$ is cost/cent, or *cost change per cent shifted*. Therefore, to obtain an update amount in cents for (8), the step size μ has the unit cent^2/cost and can be interpreted as the pitch shift applied per frame when the cost gradient equals one. For example, with $\mu = 50$ and a frame length of 20 ms (a frame rate of 50 Hz, as in all examples presented here), a sustained unit gradient would result in a pitch shift rate of $50 \times 50 = 2500$ cents per second. However, the actual maximum gradient of C_{t} is $\pi K/1200 = \pi/100 \approx 0.031$, so that even in the most extreme case, $\mu = 50$ yields a maximum shift rate of approximately 78.5 cents per second.[3]

In the following Sections IV-B to IV-E, we introduce the hyperparameters of the algorithm and discuss their effects, illustrated also in Fig. 5 using the S voice of our example cadence. For each varied hyperparameter, Fig. 5 shows the resulting pitch shift curve $p(n)$ as well as the 12-TET offset of the target voice overlaid on the cost landscape, before (orange) and after (shades of gray) applying $p(n)$. To keep these illustrations simple, we set $w = 1$ in (1), using the cost C_{t} only. The harmonic cost C_{h} and the extension to a multi-voice setting are then introduced in Section IV-F.

B. Number of Iterations

As a first hyperparameter for controlling the amount of adaptation, we vary the number of gradient descent iterations L, using an initial value of $p^{(0)}(n) = 0$ and a fixed step size of $\mu = 50$. The three shades of gray in Fig. 5(a) show the resulting $p(n)$ for $L = 1$, $L = 10$, and $L = 100$, from dark to light. With

[3]For C_{h}, the maximum gradient is less straightforward to derive analytically, but is expected to be smaller than that of C_{t}.

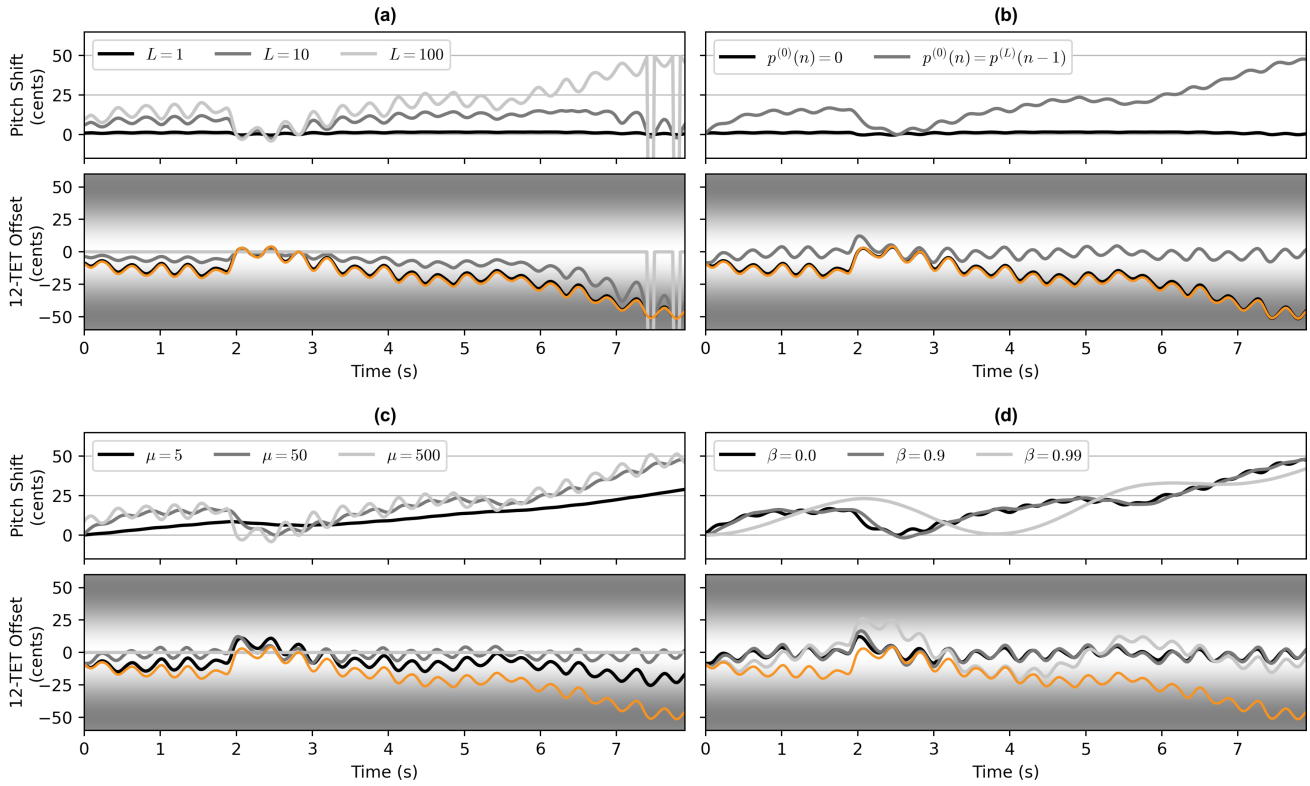

Fig. 5. Examples for intonation adaptation with different hyperparameter settings using only the tonal cost ($w = 1$) and the S voice from the example cadence. Each panel (a-d) shows the resulting pitch shift curves $p(n)$ (top) and the original and adapted 12-TET offsets overlaid on the cost landscape (bottom). (a) Varying the number of iterations $L = 1$ (dark gray), $L = 10$ (medium gray), and $L = 100$ (light gray), with $p^{(0)}(n) = 0$ and $\mu = 50$. (b) Varying the initial conditions $p^{(0)}(n) = 0$ (dark gray), and $p^{(0)}(n) = p^{(L)}(n-1)$ (medium gray), with $L = 1$ and $\mu = 50$. (c) Varying the step size $\mu = 5$ (dark gray), $\mu = 50$ (medium gray), and $\mu = 500$ (light gray), with $L = 1$ and $p^{(0)}(n) = p^{(L)}(n-1)$. (d) Varying the momentum parameter $\beta = 0.0$ (dark gray), $\beta = 0.9$ (medium gray), $\beta = 0.99$ (light gray), with $L = 1$, $p^{(0)}(n) = p^{(L)}(n-1)$, and $\mu = 50$.

$L = 1$, the adaptation has virtually no effect, as a single gradient step is insufficient to meaningfully reduce the cost with the given step size. At the other extreme, $L = 100$ successfully counteracts the global intonation drift, but also removes local pitch variations such as vibrato, and may push the adapted pitch towards a neighboring note whenever the original 12-TET offset exceeds 50 cents (cf. the two jumps in $p(n)$ after 7 seconds), independently of the temporal context. An intermediate value of $L = 10$ offers a compromise, but reveals that varying L alone does not allow for independent control over local and global adaptation behavior.

C. Initial Value

A key modification for incorporating temporal context is to initialize each frame with the previous result, setting

$$p^{(0)}(n) = p^{(L)}(n-1). \tag{10}$$

Fig. 5(b) compares both initial conditions with $L = 1$ and all other hyperparameters unchanged. Initializing from the previous frame (lighter gray) yields a pitch shift curve that effectively compensates for the global drift while preserving most of the local vibrato. It also allows for 12-TET offsets above 50 cents without snapping to a neighboring note, since the temporal context anchors the adaptation to the preceding frame. This reflects the temporal consistency of intonation decisions made by performers, who naturally only slowly update their frame of reference instead of reassessing independently at each moment in time.

D. Step Size

The step size μ directly controls the strength of the adaptation per frame. As shown in Fig. 5(c), a small value of $\mu = 5$ (darkest gray) yields a gentle adaptation that preserves most local fluctuations, whereas a large value of $\mu = 500$ (lightest gray) counteracts even fine-grained variations such as vibrato, similar to the effect of a large L.

E. Momentum

While μ offers intuitive control over the overall adaptation strength, it does not allow for precise and independent control over the reaction to global versus local pitch variations, motivating the introduction of an additional mechanism to address this trade-off. To do this, we introduce a *momentum* parameter $\beta \in [0, 1]$, as commonly used in gradient-based optimization [42]. To do this, we modify the update rule from (8) to

$$p^{(\ell)}(n) = p^{(\ell-1)}(n) - \mu b^{(\ell)}(n), \tag{11}$$

where

$$b^{(\ell)}(n) = \beta b^{(\ell-1)}(n) + (1-\beta)g^{(\ell)}(n) \tag{12}$$

is an exponentially weighted moving average of the gradients across iterations, with $b^{(0)}(n) = b^{(L)}(n-1)$, analogous to Section IV-C. Setting $\beta = 0$ reduces to the original update rule (8), while $\beta = 1$ freezes the update at $b^{(0)}(n)$, preventing any adaptation.

The results for $\mu = 50$ and varying β are shown in Fig. 5(d). The momentum acts as a low-pass filter on the gradient sequence, effectively suppressing the influence of rapid local fluctuations while allowing the algorithm to react to slower global trends. However, since momentum introduces a group delay proportional to $1/\log(\beta)$, analogous to a low-pass filter, choosing β too large can slow the reaction to pitch changes and an oscillating "filter resonance", as visible for $\beta = 0.99$ (light gray pitch shift curve in Fig. 5(d)).

F. Multi-Voice Adaptation

Up to this point, we have considered a single target voice adapted against a fixed accompaniment, with iterations over time frames n and optimization steps ℓ. With a set of monophonic voices, the gradient-based framework naturally extends to the joint adaptation of multiple voices, where each voice alternates between the role of target voice and accompaniment. This is particularly relevant for modeling the dynamic tuning interactions in ensembles such as string quartets, wind ensembles, or choirs. Since this kind of multi-voice adaptation is only meaningful when the cost reflects the interaction between voices, we set $w = 0.33$ in the following example, combining both cost terms to encourage JI-like intonation while maintaining tonal stability.

For a formal definition, let $\mathcal{V}$ denote the set of monophonic voices in the performance, e.g., $\mathcal{V} = \{\mathtt{S}, \mathtt{A}, \mathtt{T}, \mathtt{B}\}$ for a quartet. Each voice $v \in \mathcal{V}$ can take the role of the target voice, while the accompaniment is formed by the combination of the remaining voices. We denote the individual pitch shift curves by $p_v(n)$ and introduce a third iteration over $\mathcal{V}$ for each time frame n. Crucially, the cost function for each voice is evaluated with the pitch shifts $p_v(n-1)$ of all voices from the previous frame applied to $\mathcal{P}^{\text{tgt}}$ and $\mathcal{P}^{\text{acc}}$. This mirrors real ensemble performance, where intonation decisions in the current moment are informed by what happened in the preceding one. Since all adaptations for a given frame n are solely based on $p_v(n-1)$, the order in which individual voices are processed is irrelevant.

The effect of multi-voice adaptation is illustrated in Fig. 6. Fig. 6(a) shows the resulting pitch shift curves $p_v(n)$, and Fig. 6(b) to (e) show the 12-TET offsets of each voice overlaid on the cost landscape combining C_{t} and C_{h}. The algorithm successfully counteracts the global drift in all voices while preserving vibrato and retaining chord-specific intonation tendencies, such as the lowered major third in S during the first chord. An interesting edge case is visible in the B voice (Fig. 6(e)) during the third chord, where the cost landscape shows a tonally unstable region, yet the algorithm maintains tonal stability as opposing gradients across voices tend to cancel out over time.

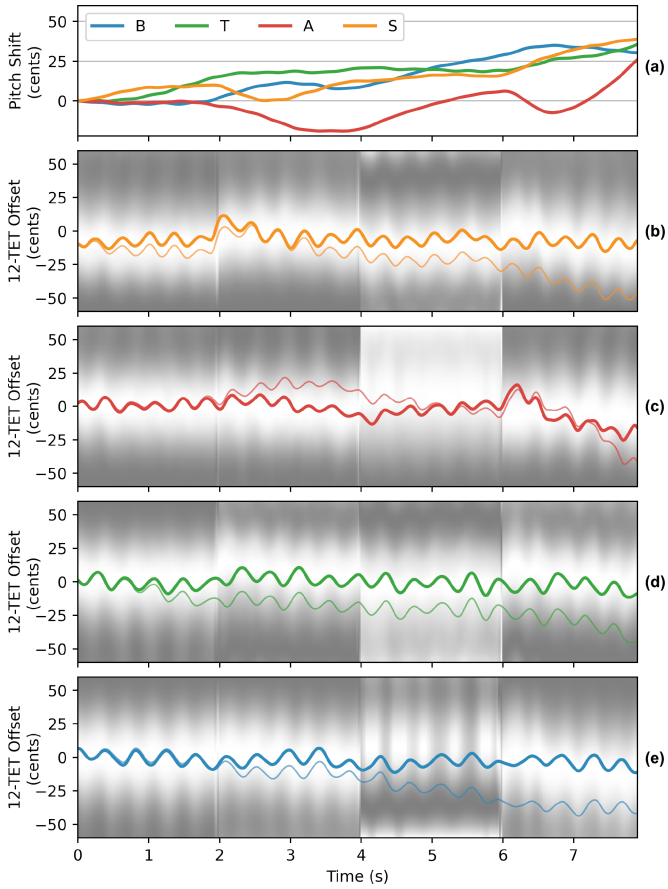

Fig. 6. Example for multi-voice adaptation in the example cadence from Fig. 1, using hyperparameters $w = 0.33$, $L = 1$, $\mu = 50$, and $\beta = 0.9$. (a) Pitch shift curves $p_v(n)$ for each voice (same as Fig. 1(c)). (b) 12-TET offset of the S voice before (thin line) and after adaptation (thick line) overlaid on the cost landscape. (c) Like (b), but for the A voice. (d) Like (b), but for the T voice. (e) Like (b), but for the B voice.

V. Experiments

The proposed method offers three main advantages over existing approaches. First, the combination of a tonal and harmonic cost term within a unified cost-based framework yields better stability in realistic musical contexts compared to methods based solely on sensory dissonance minimization [29], while allowing for flexible control between targeting 12-TET or JI-like intonation. Second, the gradient-based multi-voice adaptation with temporal consistency can handle large intonation drifts that grid-based methods cannot compensate for without score information or manual intervention. Third, the hyperparameters L, μ, and β provide effective control over the adaptation of global and/or local intonation deviations.

To demonstrate these advantages, we conduct three experiments with real-world music signals, using quartet recordings of three different ensemble types. We evaluate the properties of the multi-voice adaptation quantitatively, analyzing the resulting pitch shift curves and cost landscapes in terms of the suitability for 12-TET and JI-like adaptation (Section V-B), the compensation of intonation drift without manual intervention (Section V-C), and the trade-offs between global and local intonation adaptation (Section V-D).

Additionally, we report on a small-scale listening test (Section V-E) that evaluates listener preferences for the weighting of the cost terms, in comparison to a commercial baseline. It is important to note that we do not aim to establish the perceptual superiority of our method over existing ones, as listener preferences are influenced by many factors beyond the adaptation algorithm itself, including the quality of the pitch shifting method, the musical context, and the discrimination ability of individual listeners. Rather, the subjective evaluation should be understood as a verification that the proposed method can produce results comparable to industry-standard approaches, while offering the additional flexibility of the cost-based formulation and gradient-based adaptation.

A. Datasets

Our method requires monophonic input tracks, from which the signal representations f_0^{tgt} and $\mathcal{P}$ can be calculated. While this is a common way to record popular music (e.g., through *studio overdubbing*), advances in musical source separation (see, e.g., [43]) or specialized recording techniques [44] could also extend the applicability to music that was recorded in a shared acoustic space. For the experiments below, we create a test set of recordings from three datasets which already provide individual tracks with little or no crosstalk between voices. All signals are resampled to a sampling rate of 44100 Hz. F0 annotations are interpolated to a common frame rate of 882 samples (20 ms) before calculating $\mathcal{P}$ for each track and frame as described in Section III-A.

ChoraleBricks (CB) [45] provides recordings of Baroque four-part chorales, where each voice is played on various wind instruments and recorded in isolation using a conductor video for temporal synchronization. Aligned chord and F0 annotations are available in the dataset. We use the chorale *Befiehl Du Deine Wege*, played on brass instruments trumpet (S), fluegelhorn (A), and baritone horn (T and B).

Dagstuhl ChoirSet (DC) [46] is a multi-track dataset of a cappella choral music, performed by amateur singers. In addition to spot microphone recordings of each individual voice with little crosstalk, the dataset provides aligned score and F0 annotations. We use a recording of the four-voice motet *Locus Iste* by Anton Bruckner performed with one singer per voice (take `QuartetB T4` from the dataset).

Virtuoso Strings (VS) [47] contains multi-track recordings of string quartet performances, using spot microphones with little crosstalk between the simultaneously playing instruments. We use a single recording of an excerpt from the String Quartet Op. 74 No. 1 by Joseph Haydn (take `Q-JH-VN1-NR 10` from the dataset) and create F0 annotations for each track using the SWIPE implementation from [48].

B. 12-TETand JI Adaptation

The mathematical framework of cost-based adaptation enables the combined application of grid-based approaches (like our tonal cost C_{t}) and multi-voice interactions (like our harmonic cost C_{h}). Using the CB performance as an example scenario, we demonstrate how the combination of the two terms stabilizes the optimization while still enabling JI-like intonation without requiring score information. We run the proposed algorithm with hyperparameter settings $\mu = 50$, $\beta = 0.0$, $L = 1$, and $p^{(0)}(n) = p^{(L)}(n-1)$, varying the cost weighting w. To evaluate the adaptation, we calculate nominal target pitches for 12-TET and JI according to the time-aligned score and respective chord degree, which allows us to measure the deviation of adapted tones from this hypothetical target pitch. Results are shown in Table I. For each note (with a segmentation based on the aligned annotations), we compute the median F0 of all original and adapted voices and report its mean absolute deviation from the nominal 12-TET (*Median Dev. 12-TET*) and JI (*Median Dev. JI*) frequencies, as well as the range (the difference between the maximum and minimum) of the F0 within each note. All values are given in cents.

TABLE I
ADAPTATION RESULTS FOR THE CB PERFORMANCE WITH DIFFERENT COST WEIGHTINGS. ALL VALUES GIVEN IN CENTS. FOR EXPLANATION, SEE SECTION V-B.

Configuration	Median Dev. 12-TET	Median Dev. JI	Range
Original	9.7 ± 7.0	10.6 ± 7.8	24.1 ± 14.9
$w = 1$ (C_{t} only)	$\mathbf{4.5 \pm 6.0}$	6.9 ± 6.6	23.8 ± 15.9
$w = 0.33$	5.7 ± 5.0	$\mathbf{5.6 \pm 5.1}$	23.6 ± 14.9
$w = 0$ (C_{h} only)	12.5 ± 12.9	11.8 ± 12.8	25.6 ± 16.3

The closest adaptation to 12-TET is achieved with $w = 1$, reducing the mean absolute deviation of the note-wise median F0 from 9.7 to 4.5 cents. Using both cost terms ($w = 0.33$) moves the median F0 closer to JI, reducing the deviation from 10.6 to 5.6 cents. Using only C_{h} ($w = 0.00$), which corresponds to the approach from [29], results in median F0 deviations exceeding those of the original. This confirms that the tonal cost term is essential for a stable adaptation. The remaining deviations in the adapted voices can partly be attributed to the delayed response in $p_v(n)$ to large 12-TET offsets in very short notes. Across all settings, the F0 range within notes is largely maintained, indicating that local F0 variations are preserved with the chosen hyperparameters. The trade-off between absolute accuracy and local variation is discussed further in Section V-D.

C. Intonation Drift Compensation

Especially in vocal ensemble recordings, compensating for intonation drift can be an important post-production step, for example when combining multiple takes into a coherent performance. While the drift in our synthetic cadence occurred over a short time span, drift in real performances is often more subtle and non-linear, with physiological [4] or musical [9] origins. The DC performance used in this application example exhibits a global drift of over 100 cents, with the ensemble collectively singing around a semitone lower by the end of the recording. Compensating for this drift while preserving local intonation is not possible with grid-based methods without score information, since they would typically adapt towards the nearest semitone. With suitable hyperparameter settings, our method can account for the drift without manual intervention.

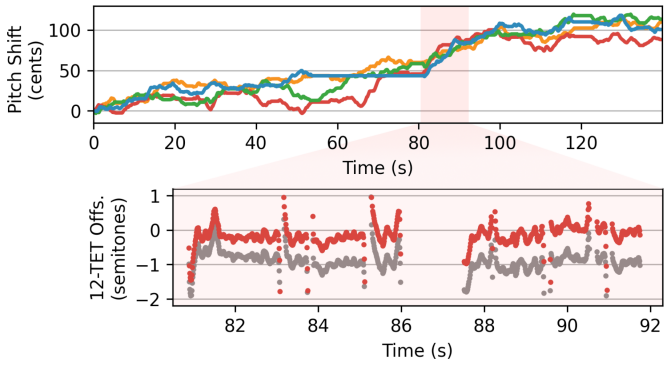

Fig. 7. Adapting global intonation drift in DC with $w = 1$, $\mu = 5$, $\beta = 0.9$, $L = 1$. The upper plot shows the pitch shift curves $p_v(n)$ over the entire piece (S: orange, A: red, T: green, B: blue). The lower plot shows the 12-TET offset of the A voice for an excerpt (gray: original, red: adapted).

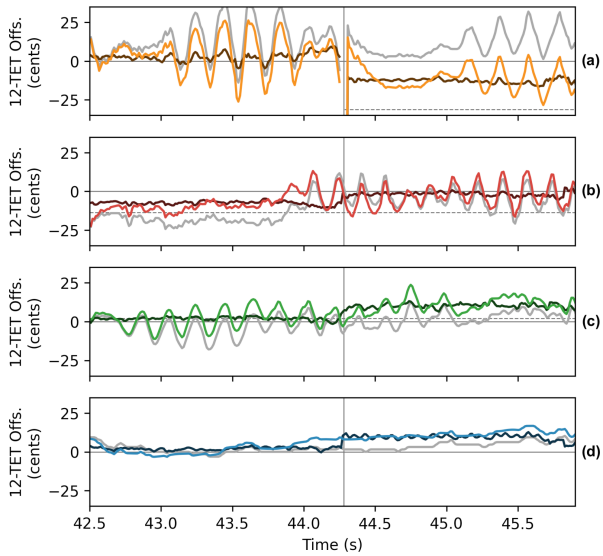

Fig. 8. Global vs. local intonation adaptation in an excerpt from VS using $w = 0.25$ (larger emphasis on C_{h}). Each plot shows the original F0 offset (gray dots), the F0 offset of the voices adapted with $\mu = 50$, $\beta = 0.9$, $L = 1$ (lighter color) and $\mu = 50$, $\beta = 0.0$, $L = 100$ (darker color). The vertical line indicates a change of the chord from D:maj to D:7. The horizontal dashed line indicates JI intonation (with a 7/4 minor seventh in the S voice). (a) S voice, (b) A voice, (c) T voice, (d) B voice.

To compensate for drift while preserving local fluctuations, we choose a small step size $\mu = 5$ and use momentum ($\beta = 0.9$) to further smooth the pitch shift curves $p_v(n)$. The resulting $p_v(n)$ for the entire performance are visualized in Fig. 7. It can be seen that while $p_v(n)$ behaves similarly across individual voices, the drift is not linear over the course of the performance. As indicated by the excerpt from the A voice in the lower plot of Fig. 7, most local intonation phenomena (such as vibrato and the relative tuning of individual notes) are retained in the adapted version (shown in red).

D. Global and Local Adaptation

The hyperparameters L, μ, and β provide an interpretable way to balance between global and local intonation adaptation. To illustrate their effect in a real-world example, we fix $w = 0.25$ to target JI-like intonation and compare two contrasting configurations in an excerpt of the VS performance. The *global setting* ($\mu = 50$, $\beta = 0.9$, $L = 1$) shifts the median F0 of each note toward cost-minimizing intervals while preserving vibrato and other local fluctuations. The *local setting* ($\mu = 50$, $\beta = 0.0$, $L = 100$) applies multiple gradient steps per frame to approach a local minimum, effectively suppressing local pitch fluctuations such as vibrato.

Results are shown in Fig. 8. For each voice, we visualize the 12-TET offset of the original (in gray), after adaptation with the global setting (in lighter color corresponding to the respective voice) and with the local setting (darker color). Both settings yield similar per-note median F0 values, while the local setting does not preserve any of the vibrato. In the first chord, between 42.5 and 44.3 seconds, the median F0 lies close to the respective JI target pitch (shown as horizontal dashed lines and clearly visible for the A voice playing a major third) due to the strong weighting of the harmonic cost. In the second chord of this example, the S voice plays a minor seventh in a D:7 chord. In this case, the deviation from 12-TET is distributed among the voices, with the adapted A, T and B voices considerably above their nominal JI target pitches.

This example also exposes two inherent limitations of the proposed approach. First, the harmonic cost C_{h} does not allow the user to select which JI ratio is targeted. Depending on the musical context, there are several options for tuning the minor seventh interval in the S voice relative to the root, including the harmonic seventh with an interval ratio of $7/4$ (or -31 cents relative to 12-TET), as well as ratios of $16/9$ (-4 cents) or $9/5$ ($+8$ cents). The harmonic cost attains a minimum near -31 cents, consistent with both the downward shift of the S voice and the upward shifts of the remaining voices. This results in a strong deviation from 12-TET intervals that may not be musically desirable and is an inherent consequence of using sensory dissonance minimization as a proxy for JI-like intonation, as the location of the local minimum depends on the voice timbres and may not coincide with the artistic intent. Second, preserving local intonation variations comes at the cost of slower adaptation at chord transitions, a trade-off that is clearly visible at the chord change at 44.3 seconds, where the local setting responds more rapidly.

E. Listening Test

To complement the methodological discussion above, we conducted a small-scale listening test with expert listeners to examine whether combining the two cost terms is advantageous compared to optimizing only tonal or only harmonic aspects, and to anchor the perceptual quality in comparison to a strong commercial baseline. Specifically, we performed a paired comparison test with five different conditions.

OR: The original, unaltered excerpt.

ME: Each voice was separately processed with the commercial pitch correction software *Melodyne Essential v5.4.2* [15], using automatic F0 estimation and segmentation with the "Melodic"

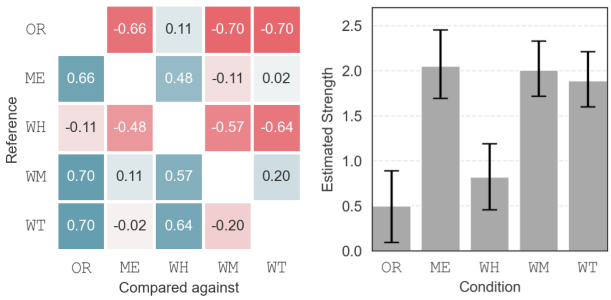

Fig. 9. Results of listening test with 11 participants. OR: Original, ME: Melodyne, WH ours with $w = 0$, WM: ours with $w = 0.33$, WT: ours with $w = 1$. **(left)** Average rating for each paired comparison. **(right)** Relative strengths for each condition estimated using a CLMM model [50], showing mean and 95% confidence interval obtained by 1000× bootstrapping.

algorithm and subsequently running the "Correct Pitch" algorithm with both "Pitch Center" and "Pitch Drift" set to 100%. By using only automatic adaptations, we ensure comparability with our method.

WH: Our method, using only the harmonic cost ($w = 0$). This can be seen as a version of the method proposed in [29].

WM: Our method, mixing both cost terms ($w = 0.33$).

WT: Our method, using only the tonal cost ($w = 1$).

For easier comparisons, we kept all other hyperparameters fixed for our approach with $\mu = 50$, $\beta = 0.9$, and $L = 1$. No condition requires manual intervention by a sound engineer. As test items, we used excerpts of four to eight seconds length from the application scenarios DC, VS, and CB, as well as the synthetic example cadence. For applying the pitch shift curves $p_v(n)$ to the audio signals, we used the time-varying pitch shifting from [41] for WH, WM, and WT, and the proprietary pitch shifting of Melodyne for ME.

In an online experiment using *WebMUSHRA* [49], participants were presented with paired comparisons in random order, with four items and five conditions leading to $4 \times ((4 \times 5)/2) = 40$ comparisons in total. In each comparison, the two stimuli should be rated in terms of the perceived intonation quality and naturalness on a five-point Likert scale with values '2' (stimulus A is much better), '1' (A is a bit better), '0' (no preference), '-1' (B is a bit better), and '-2' (B is much better). The signals were presented in mono over headphones. Of 11 participants in total (median age 33 years), 9 reported "a lot" of or "expert"-level experience in judging musical performances in a post-test questionnaire and all of them reported "a lot" of or "expert"-level experience as musicians. All listeners participated voluntarily and consented to the processing of their responses in an anonymized way.

The left side of Fig. 9 shows the mean rating for each paired comparison, averaged over participants and test items. It can be observed that WM, WT, and ME achieve the largest subjective improvement over the OR condition, with an average rating of 0.70, 0.70, and 0.66, respectively. WH, on the other hand, does not seem to lead to favorable intonation and is rated worse than all other conditions on average. In direct comparison, WM is slightly preferred over WT and ME with an average rating of 0.20 and 0.11, respectively, while there is almost no difference between WT and ME (average rating of -0.02 in favor of ME). We additionally analyzed the occurrence of circular judgments (such as preferring A over B, B over C, and C over A) which were observed in 16 of 44 possible cases (11 listeners and 4 test items). This may indicate limited perceptual discriminability between the test items [51] and is consistent with post-test reports from some listeners, who indicated difficulty in forming clear preferences for some comparisons.

For further analysis, we employ a cumulative link mixed model (CLMM) [50], [52] to estimate a *relative strength* of each condition, indicating the likelihood that a certain condition is preferred over the others. In the CLMM, we take individual differences between listeners and test items into account by modeling both as random effects. We did not observe issues with model convergence due to circular or otherwise inconsistent ratings. By applying bootstrapping with 1000 resamplings of the listener responses, we estimate mean and 95% confidence intervals of the estimated strengths which are displayed as a bar plot with error bars on the right side of Fig. 9. While OR and WH are clearly separated from the other conditions, ME, WM, and WT have overlapping confidence intervals and a mean strength of 2.06, 2.01, and 1.89, respectively.

ME and WT share the same general objective (aligning the pitch with the 12-TET grid), so that they should theoretically yield similar results. However, several other aspects could influence the observed difference in perceived intonation quality and naturalness, including the behavior of the pitch shift curves at note onsets (due to the previous segmentation step, ME allows for faster adaptation) and differences in the quality of the pitch shift algorithms. Notably, WM uses the same hyperparameters (except w) and pitch shift algorithm as WT, yet is closer to ME in estimated strength and is even slightly preferred in the mean ratings, hinting at an overall positive effect of combining the two cost terms. WH (using only harmonic cost, as proposed in [29]) does not yield favorable results, possibly due to limitations in the accuracy of gradients. While this small-scale listening experiment only provides tentative insights, the results suggest that combining cost terms that model different intonation aspects can yield results that are preferred over tonal or harmonic adaptation individually.

VI. Conclusions and Outlook

In this paper, we introduced a method for intonation adaptation via gradient descent. Based on a differentiable cost measure that models different aspects of favorable intonation in a multi-voice context, our method can handle intonation drift in the input and allows for a flexible configuration of hyperparameters to accommodate different artistic preferences between 12-TET and JI-like intonation. In future work, the approach could be extended by modeling additional aspects such as melodic intonation in the cost measure. Furthermore, larger-scale subjective experiments with real-world music signals are needed to better understand in which situations different intonation influences are more or less perceptually relevant.

APPENDIX

A. Differentiability of (5)

Proposition 1: 1 The function $d_{\mathrm{h}}(f, f')$ (5) is differentiable with respect to f with $f, f', d_{\max} \in \mathbb{R}_{>0}$.

Proof: We rewrite $d_{\mathrm{h}}(f, f') = \phi(\psi(f, f'))$ as the composition of two functions

$$\psi(f, f') = \frac{\log_2(f/f')}{d_{\max}} \tag{13}$$

with $f, f', d_{\max} \in \mathbb{R}_{>0}$ and

$$\phi(x) = \begin{cases} \exp\left(-\ln^2(|x|)\right) & x \neq 0 \\ 0 & x = 0 \end{cases} \tag{14}$$

with $x \in \mathbb{R}$.

$\psi(f, f')$ is a composition of functions that are differentiable in its entire domain (including $f = f'$), so it is itself differentiable. The same is easily verified for $\phi(x)$ with $x \neq 0$, while for $x = 0$, we need to show that the limit

$$\lim_{h \to 0} \frac{\phi(0 + h) - \phi(0)}{h} \tag{15}$$

exists from both sides. For this, we rewrite the piece-wise definition as

$$\phi(x) = \begin{cases} \exp\left(-\ln^2(-x)\right) & x < 0 \\ 0 & x = 0 \\ \exp\left(-\ln^2(x)\right) & x > 0 \end{cases} \tag{16}$$

and obtain

$$\begin{aligned} &\lim_{h \to 0^+} \frac{\exp\left(-\ln^2(h)\right) - 0}{h} \\ &= \lim_{h \to 0^+} \frac{\exp\left(-\ln^2(h)\right)}{\exp(\ln(h))} \\ &= \lim_{h \to 0^+} \exp\left(-\ln^2(h) - \ln(h)\right) \\ &= \lim_{h \to 0^+} \exp\left(\underbrace{-\ln(h)}_{\to \infty} \underbrace{(\ln(h) + 1)}_{\to -\infty}\right) = 0, \end{aligned} \tag{17}$$

$$\begin{aligned} &\lim_{h \to 0^-} \frac{\exp\left(-\ln^2(-h)\right) - 0}{h} \\ &= \lim_{h \to 0^+} \frac{\exp\left(-\ln^2(h)\right) - 0}{-h} = 0. \end{aligned} \tag{18}$$

$d_{\mathrm{h}}(f, f')$, as the composition of two differentiable functions $\phi(x)$ and $\psi(f, f')$, is therefore also differentiable. □

ACKNOWLEDGMENT

The authors thank the anonymous reviewers for their valuable comments. The International Audio Laboratories Erlangen are a joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) and Fraunhofer Institute for Integrated Circuits IIS.

REFERENCES

[1] J. M. Barbour, *Tuning and Temperament: A Historical Survey*. East Lansing, MI, USA: Michigan State College Press, 1951.
[2] F. Havrøy, "'You cannot just say: 'I am singing the right note". Discussing intonation issues with Neue Vocalsolisten Stuttgart," *Music Pract.*, vol. 1, 2013.
[3] T. J. Cuffman, "A practical introduction to just intonation through string quartet playing," DMA dissertation, Univ. Iowa, Iowa City, IA, USA, 2016.
[4] P.-G. Alldahl, *Choral Intonation*. Hägersten, Stockholm: Gehrmans Musikförlag, 2008.
[5] S. Applebaum and T. Lindsay, *The Art and Science of String Performance*. Sherman Oaks, CA, USA: Alfred Publishing Company, 1986.
[6] E. Prame, "Vibrato extent and intonation in professional western lyric singing," *J. Acoustical Soc. Amer.*, vol. 102, no. 1, pp. 616–621, 1997.
[7] J. M. Geringer, R. B. MacLeod, C. K. Madsen, and J. Napoles, "Perception of melodic intonation in performances with and without vibrato," *Psychol. Music*, vol. 43, no. 5, pp. 675–685, 2015.
[8] C. Harte, M. B. Sandler, S. Abdallah, and E. Gómez, "Symbolic representation of musical chords: A proposed syntax for text annotations," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, London, U.K., 2005, pp. 66–71.
[9] D. M. Howard, "Intonation drift in a capella soprano, alto, tenor, bass quartet singing with key modulation," *J. Voice*, vol. 21, no. 3, pp. 300–315, 2007.
[10] R. Seaton, D. Sharp, and D. Pim, "Pitch drift in a capella choral singing: The outcomes from an international survey," in *Proc. Inst. Acoust.*, 2014, vol. 36, no. 3, pp. 312–319.
[11] S. Schwär, S. Rosenzweig, and M. Müller, "A differentiable cost measure for intonation processing in polyphonic music," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, 2021, pp. 626–633.
[12] J. Berezovsky, "The structure of musical harmony as an ordered phase of sound: A statistical mechanics approach to music theory," *Sci. Adv.*, vol. 5, no. 5, 2019, Art. no. eaav8490.
[13] W. A. Sethares, *Tuning, Timbre, Spectrum, Scale*. Berlin, Germany: Springer, 1998.
[14] P. M. C. Harrison and M. T. Pearce, "Simultaneous consonance in music perception and composition," *Psychol. Rev.*, vol. 127, pp. 216–244, Mar. 2020.
[15] Celemony, "Melodyne," Accessed: Mar. 17, 2026. [Online]. Available: https://www.celemony.com/en/melodyne/what-is-melodyne
[16] Antares, "Auto-tune," Accessed: Mar. 17, 2026. [Online]. Available: https://www.antarestech.com/products/auto-tune/pro
[17] Apple, "Logic pro x and flex pitch," Accessed: Mar. 17, 2026. [Online]. Available: https://www.apple.com/logic-pro/
[18] D. Code, "Groven.Max: An adaptive tuning system for MIDI pianos," *Comput. Music J.*, vol. 26, pp. 50–61, 2002.
[19] W. Mohrlock, "Hermode tuning," Accessed: Mar. 17, 2026. [Online]. Available: http://www.hermode.com/index_en.html
[20] K. Stange, C. Wick, and H. Hinrichsen, "Playing music in just intonation: A dynamically adaptive tuning scheme," *Comput. Music J.*, vol. 42, no. 3, pp. 47–62, 2018.
[21] Y.-J. Luo, M.-T. Chen, T.-S. Chi, and L. Su, "Singing voice correction using canonical time warping," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Calgary, AB, Canada, 2018, pp. 156–160.
[22] S. Yong and J. Nam, "Singing expression transfer from one voice to another for a given song," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Calgary, AB, Canada, 2018, pp. 151–155.
[23] X. Zhuang, H. Yu, W. Zhao, T. Jiang, and P. Hu, "KaraTuner: Towards end-to-end natural pitch correction for singing voice in karaoke," in *Proc. Interspeech*, 2022, pp. 4262–4266, doi: 10.21437/Interspeech.2022-939.
[24] J. Hai and M. Elhilali, "Diff-pitcher: Diffusion-based singing voice pitch correction," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, 2023, pp. 1–5.
[25] S. Wager et al., "Intonation: A dataset of quality vocal performances refined by spectral clustering on pitch congruence," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Brighton, U.K., 2019, pp. 476–480.
[26] S. Wager, G. Tzanetakis, C. Wang, and M. Kim, "Deep autotuner: A pitch correcting network for singing performances," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Barcelona, Spain, 2020, pp. 246–250.
[27] H. Hyodo, S. Takamichi, T. Nakamura, J. Koguchi, and H. Saruwatari, "DNNbased ensemble singing voice synthesis with interactions between singers," in *Proc. IEEE Spoken Lang. Technol. Workshop*, Macao, China, 2024, pp. 660–667.
[28] W. A. Sethares, "Local consonance and the relationship between timbre and scale," *J. Acoust. Soc. Amer.*, vol. 94, no. 3, pp. 1218–1228, 1993.

[29] W. Sethares, "Adaptive tunings for musical scales," *J. Acoust. Soc. Amer.*, vol. 96, pp. 10–18, 1994.
[30] J. Villegas and M. Cohen, "Roughness minimization through automatic intonation adjustments," *J. New Music Res.*, vol. 39, pp. 75–92, 2010.
[31] R. Parncutt and G. Hair, "A psychocultural theory of musical interval: Bye bye pythagoras," *Music Percep.*, vol. 35, no. 4, pp. 475–501, 2018.
[32] S. Schwär, S. Balke, and M. Müller, "Measuring sensory dissonance in multi-track music recordings: A case study with wind quartets," in *Proc. Int. Soc. Music Inf. Retrieval Conf.*, Daejeon, South Korea, 2025, pp. 117–124.
[33] 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 *Proc. Int. Soc. Music Inf. Retrieval Conf.*, Delft, The Netherlands, 2019, pp. 276–283.
[34] M. Mauch, K. Frieler, and S. Dixon, "Intonation in unaccompanied singing: Accuracy, drift, and a model of reference pitch memory," *J. Acoust. Soc. Amer.*, vol. 136, no. 1, pp. 401–411, 2014.
[35] T. Nakano, M. Goto, and Y. Hiraga, "An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, Pittsburgh, PA, USA, 2006, pp. 1706–1709.
[36] J. Tenney, *A History of 'Consonance' and 'Dissonance'*. New York, NY, USA: Excelsior Music Publishing Company, 1988.
[37] R. Plomp and W. J. M. Levelt, "Tonal consonance and critical bandwidth," *J. Acoust. Soc. Amer.*, vol. 38, no. 4, pp. 548–560, 1965.
[38] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hear. Res.*, vol. 47, no. 1-2, pp. 103–138, 1990.
[39] E. Zwicker, "Subdivision of the audible frequency range into critical bands (Frequenzgruppen)," *J. Acoust. Soc. Amer.*, vol. 33, no. 2, 1961, Art. no. 248.
[40] R. Marjieh, P. M. C. Harrison, H. Lee, F. Deligiannaki, and N. Jacoby, "Timbral effects on consonance disentangle psychoacoustic mechanisms and suggest perceptual origins for musical scales," *Nature Commun.*, vol. 15, 2024, Art. no. 1482.
[41] S. Rosenzweig, S. Schwär, J. Driedger, and M. Müller, "Adaptive pitch-shifting with applications to intonation adjustment in a cappella recordings," in *Proc. Int. Conf. Digit. Audio Effects*, Vienna, Austria, 2021, pp. 121–128.
[42] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
[43] G. Richard, P. Chouteau, and B. Torres, "A fully differentiable model for unsupervised singing voice separation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Seoul, Republic of Korea, 2024, pp. 946–950.
[44] S. Schwär, M. Krause, M. Fast, S. Rosenzweig, F. Scherbaum, and M. Müller, "A dataset of larynx microphone recordings for singing voice reconstruction," *Trans. Int. Soc. Music Inf. Retrieval*, vol. 7, no. 1, pp. 30–43, 2024.
[45] S. Balke, A. Berndt, and M. Müller, "ChoraleBricks: A modular multitrack dataset for wind music research," *Trans. Int. Soc. Music Inf. Retrieval*, vol. 8, no. 1, pp. 39–54, 2025.
[46] S. Rosenzweig, H. Cuesta, C. Weiß, F. Scherbaum, E. Gómez, and M. Müller, "Dagstuhl ChoirSet: A multitrack dataset for MIR research on choral singing," *Trans. Int. Soc. Music Inf. Retrieval*, vol. 3, no. 1, pp. 98–110, 2020.
[47] M. Tomczak, M. S. Li, and M. D. Luca, "Virtuoso strings: A dataset of string ensemble recordings and onset annotations for timing analysis," in *Proc. Late-Breaking Demos Int. Soc. Music Inf. Retrieval Conf.*, Milano, Italy, 2023.
[48] S. Rosenzweig, S. Schwär, and M. Müller, "libf0: A Python library for fundamental frequency estimation," in *Proc. Late Breaking Demos Int. Soc. Music Inf. Retrieval Conf.*, Bengaluru, India, 2022.
[49] M. Schoeffler et al., "webMUSHRA—A comprehensive framework for web-based listening tests," *J. Open Res. Softw.*, vol. 6, no. 8, 2018.
[50] J. E. Taylor, G. A. Rousselet, C. Scheepers, and S. C. Sereno, "Rating norms should be calculated from cumulative link mixed effects models," *Behav. Res. Methods*, vol. 55, pp. 2175–2196, 2023.
[51] E. Parizet, "Paired comparison listening tests and circular error rates," *Acta Acustica United Acustica*, vol. 88, pp. 594–598, 2002.
[52] R. H. B. Christensen, "ordinal—Regression models for ordinal data," 2023, R package version 2023.12-4.1. [Online]. Available: https://cran.rproject.org/web/packages/report/vignettes/cite_packages.html