# Intonation in unaccompanied singing: Accuracy, drift, and a model of reference pitch memory

Matthias Mauch,[a] Klaus Frieler,[b] and Simon Dixon
*Centre for Digital Music, Queen Mary University of London, London E1 4NS, United Kingdom*

(Received 2 September 2013; revised 30 April 2014; accepted 10 May 2014)

This paper presents a study on intonation and intonation drift in unaccompanied singing, and proposes a simple model of reference pitch memory that accounts for many of the effects observed. Singing experiments were conducted with 24 singers of varying ability under three conditions (*Normal*, *Masked*, *Imagined*). Over the duration of a recording, ∼50 s, a median absolute intonation drift of 11 cents was observed. While smaller than the median note error (19 cents), drift was significant in 22% of recordings. Drift magnitude did not correlate with other measures of singing accuracy, singing experience, or the presence of conditions tested. Furthermore, it is shown that neither a static intonation memory model nor a memoryless interval-based intonation model can account for the accuracy and drift behavior observed. The proposed causal model provides a better explanation as it treats the reference pitch as a changing latent variable.
 [http://dx.doi.org/10.1121/1.4881915]

PACS number(s): 43.75.Rs, 43.75.Bc, 43.75.Xz, 43.70.Fq [DD] Pages: 401–411

## I. INTRODUCTION

Unlike other musical instruments, the vocal apparatus is common to all human beings, and in every known human culture, people use it to make music [Brown (1991), as reproduced by Pinker (2002)]. There is good evidence that vocal music was practiced even in prehistoric human societies, and it might even have preceded language (Mithen, 2007). Yet, science is only beginning to understand the control processes involved in human singing. This paper aims to provide some insights into intonation, a parameter that is crucial to many singing styles, but has so far received little academic attention.

*Intonation* is defined as "accuracy of pitch in playing or singing" (Swannell, 1992), or "the act of singing or playing in tune" (Kennedy, 1980). Both of these definitions imply the existence of a reference pitch, which could be internal or external. We treat intonation as the signed pitch difference relative to the reference pitch, measured in semitones on an equal-tempered scale (see detailed discussion in Sec. III).

In choirs, intonation is the main reported priority in daily rehearsals (Ganschow, 2013) and the focus of guides on choral practice (e.g., Crowther, 2003). Such ensembles frequently observe a change in tuning over periods of tens of seconds or even a whole piece, a phenomenon called *intonation drift* or pitch drift (Seaton *et al.*, 2013). According to Alldahl (2006), "the problem is mainly a lowering of pitch," i.e., downward intonation drift. Seaton *et al.* (2013) offer a literature review on choral intonation drift, and their pilot survey on drift in choral singing corroborates Alldahl's observation that drift mainly occurs in the downward direction. Several scientific studies suggest that one cause for the propensity to drift is the harmonic progression (Terasawa, 2004; Howard, 2007; Devaney *et al.*, 2012; see also Sec. II).

Yet, harmonic effects cannot be the only cause for intonation drift since it also occurs in solo singing: In a study on folk song analysis, Müller *et al.* (2010), tracking the tuning on a stanza level, report that intonation drift is common in unaccompanied solo folk singing. Intriguingly, their example shows that strong rises in tuning were observed, but no further investigations are reported. Ryynänen (2004) built adaptive tuning into a note transcription system based on the observation that "non-professional singers tend to change their tuning (typically downwards) during long melodies." Here, intonation drift is treated as a nuisance factor. Dalla Bella *et al.* (2007) investigate pitch stability as one of several variables describing pitch in singing. To our knowledge, no other studies on unaccompanied solo singing exist that investigate intonation drift in its own right.

Hence, the main motivation for the present study is to improve the scientific understanding of intonation drift in unaccompanied solo singing, without additional influences of harmonic consonance or ensemble interaction. Findings drawn from observations in this simpler setting are likely to play a role in explaining drift in complex ensemble situations as well. In order to understand which mechanisms may cause drift, we study three different conditions, *Normal*, *Masked*, and *Imagined* (Sec. III).

The remainder of the paper is structured as follows. Section II discusses existing work related to singing intonation and musical memory. Section III describes our intonation experiments, including the three experimental conditions, as well as a basic outline of the analysis setup. Section IV defines and illustrates several metrics of singing accuracy and drift. In Sec. V, we show which intrinsic and external factors influence accuracy and drift. Section VI introduces a simple model of pitch reference memory, which is able to account for the intonation stability and drift we observed. Section VII provides a discussion of achievements

[a]Author to whom correspondence should be addressed. Electronic mail: matthias.mauch@eecs.qmul.ac.uk
[b]Also at: Musikwissenschaftliches Institut, HfM Franz Liszt Weimar.

and future work, and a summary of our conclusions is found in Sec. VIII.

## II. PREVIOUS WORK

Only since the advent of precise pitch analysis, in the form of the tonoscope (Seashore, 1914), has it been possible to study intonation quantitatively. Carl Seashore's *Psychology of Music* (Seashore, 1967, originally published in 1938) already featured analyses of vibrato based on this technique. Since then, less burdensome methods for pitch analysis have been devised (e.g., Schroeder, 1968; Markel, 1972; de Cheveigné and Kawahara, 2002). These methods, along with computer programs like Praat (Boersma, 2002) and the advent of fast, affordable computers, have made intonation analysis accessible to anyone with a microphone and a computer.

Recently, progress has been made on quantifying differences in intonation between singers. In the music informatics domain, singing tuition applications (e.g., Cano *et al.*, 2012) have driven the development of singing assessment methods that often focus on intonation aspects (for an overview, see Molina, 2012). In the music psychology literature, the phenomenon of so-called "poor singers" has gained some interest (e.g., Dalla Bella *et al.*, 2007; Berkowska and Dalla Bella, 2009; Dalla Bella and Berkowska, 2009; Pfordresher *et al.*, 2010). Welch (1985) proposed a theory of singing production, with special regard as to how children acquire singing skills.

Vurma and Ross (2006) investigated professional singers' ability to sing intervals and reported average standard deviations of 22 cents in interval size, and 34 cents in absolute pitch, relative to a tuning fork reference. Immediately after singing, the singers were unable to judge whether their intervals were out of tune, but after listening to a recording of their singing, their judgments were not significantly different from other expert listeners. Judgments of out-of-tune singing correlated with pitch errors, but errors of even 40 cents were not reliably judged out of tune by the majority of listeners.

Dalla Bella *et al.* (2007) compared occasional and professional singers performing a well-known melody in a free memory recall scenario. Two groups of occasional singers made errors in singing intervals of around 0.6 and 0.9 semitones on average, while professional singers' errors were only 0.3 semitones. A correlation with tempo was also observed, and a second experiment was performed, which confirmed that errors decreased significantly when the same singers sang more slowly. In a further study, Dalla Bella and Berkowska (2009) used both free recall and repetition paradigms to characterize poor singing in terms of timing accuracy, relative pitch (interval) accuracy, and absolute pitch accuracy, and found that poor singers could have deficits in any one or any combination of these attributes.

Pfordresher *et al.* (2010) distinguished the accuracy (mean deviation from a target pitch) and precision (consistency in repeated attempts to produce a pitch) of singers in order to classify "poor" singers. They found that the majority (56%) of singers were imprecise (standard deviation of pitch error $>1$ semitone), but only 13% of singers were inaccurate (absolute value of average error $>1$ semitone). It was also observed that errors were greater for the imitation task than for a recall task.

Most existing research on intonation is concerned with a fixed tuning system, but some authors have also studied intonation drift. Terasawa (2004), Howard (2007), and Devaney *et al.* (2012) investigated pitch drift in unaccompanied vocal ensembles. In such a context, physics predicts that perfect consonance conflicts with pitch stability over time. The idea goes back at least to the 16th century, when music theorist Giovanni Benedetti wrote a piece of three-part singing designed to result in various amounts of pitch drift. The evidence from the new studies for a reliably predictable effect is not entirely conclusive, partly due to small sample sizes: Devaney *et al.* (2012) reported only negligible effects on the original Benedetti composition, while Howard (2007) reported drifts roughly in line with predictions on specially composed new pieces. Dalla Bella *et al.* (2007) also measured *pitch stability* and found absolute deviations between repeated sequences of notes of 0.3 semitones in professional singers and 0.6 semitones in occasional singers.

## III. METHOD

### A. Participants

A total of 31 participants from the UK and Germany took part in the experiment. They were recruited from musicology students, office colleagues, lab members, and the choir of the Wolfson College in Cambridge, UK. Our aim is to study intonation of subjects who are not "poor" singers (Pfordresher and Brown, 2007). Hence, two participants were excluded because they produced a melody that matched "Happy Birthday" rhythmically, but not tonally (they consistently sang a different melody). A third singer had an unstable voice from which we were unable to draw suitable pitch estimates. Also excluded were four further participants, who were detected as outliers and hence classified as "poor" singers. The outlier classification was performed using multivariate outlier detection (Filzmoser *et al.*, 2005) on two singer-based metrics: mean absolute interval error (see Sec. IV C) and ratio of intervals within a semitone of the true interval. After these exclusions, 24 subjects remained in the study. The age of the participants ranged from 13 to 62 with a median of 32.5 yr (mean: 34.5). The gender ratio was imbalanced with 6 females and 18 males in the sample. The musical experience of participants was widespread. Fourteen singers considered themselves amateur musicians, nine professionals or semi-professionals, and one reported no musical background. Thirteen participants reported "a lot" of singing experience, nine some or no experience, one subject sings on a professional level, and one did not respond. Eleven subjects are still active in some choir, while eight had previous choir experience, and five have never sung in a choir (see Table I). Since we had a large share of male participants, baritone was the most common voice type with a total of 13 subjects, followed by soprano with 6 subjects.

### B. Material

Since we chose to employ a free memory recall paradigm with a variety of subjects from two different countries,

TABLE I. Self-reported musical experience.

| Musical background | | Choir experience | |
|---|---|---|---|
| None | 1 | None | 5 |
| Amateur | 14 | As a child | 3 |
| Semi-professional | 7 | No longer active | 5 |
| Professional | 2 | Still active | 11 |
| **Singing skill** | | **Singing experience** | |
| Poor | 1 | None | 3 |
| Low | 3 | Some | 6 |
| Medium | 14 | A lot | 13 |
| High | 4 | Professional | 1 |
| Very High | 2 | (no response) | 1 |

the choice fell on "Happy Birthday," probably the single best-known and most widespread song in the world. "Happy Birthday" cannot be considered a very easy song, since it contains a variety of different intervals, some of them being large jumps (see Fig. 1). The ambitus is exactly one octave using a full major scale from dominant to dominant an octave higher. The song is written in $\frac{3}{4}$ time, beginning with a two note upbeat and comprising a total of 25 notes in 4 phrases of 6, 6, 7, and 6 notes each.

### C. Procedure

Each participant sang a total of 9 renditions of "Happy Birthday" in three recordings of three runs each. Details are given below. For a particular recording each participant was asked to sing three consecutive runs of "Happy Birthday." The participants could choose the starting pitch at their own comfort in order to limit effects of regression to their comfort pitch. They were provided with a click track of moderate tempo (96 bpm) and instructed to wait four bars before beginning to sing. Subjects were instructed to sing the syllable "na" throughout. Subjects were recorded at a sample rate of 44 100 Hz with a bit depth of 32 bit [stored to 16-bit pulse code modulation (PCM)] using Audacity 2.0 running on a Windows Laptop (Microsoft Corp., Redmont, WA) or a MacBook Pro (Apple Inc., Cupertino, CA). A conventional headset (Logitech USB Headset 390, Logitech International S.A., Lausanne, Switzerland) functioned both as microphone and headphones, through which participants were provided with the click track and noise in the *Masked* condition (see below).

Three such recordings were made of each participant to test three different conditions, which differed by the way the second run of "Happy Birthday" was performed.

*Normal.* The participant sang three renditions of "Happy Birthday" as described above.

*Masked.* Pink noise at a moderate sound pressure level was applied over the headphones during the second of three renditions of "Happy Birthday."

*Imagined.* The participant was asked to remain silent during the second rendition of "Happy Birthday," while imagining singing, and resume singing at the start of the third rendition.

The reasoning behind these conditions was to study whether the absence of vocal strain reduces the tendency to drift (*Imagined* condition) and whether an impediment to auditory feedback would increase the tendency to drift (*Masked* condition). Note that the *Imagined* condition does not only remove vocal strain, but also auditory and kinesthetic feedback, as the participants can neither hear their singing nor feel singing-induced movements or the state of the vocal tract in the vicinity of the vocal folds. Anesthetizing the vocal folds has been shown to lead to a decrease in singing accuracy (Kleber *et al.*, 2013).

The sequence of conditions was held constant (in increasing order of difficulty). In each condition, subjects sang 75 notes except in the *Imagined* condition with only 50 notes. Most of the German singers sang the German version of the melody, which divides note 17 into 2 syllables at the same pitch; this extra note was disregarded in the analysis. One singer consistently missed note 19.

### D. Analysis

We use, as our reference tuning system, equal temperament (ET). We will see in Sec. IV, that for the purposes of our study, the assumption of ET does not substantially affect our results. We also assume that pitch, a perceptual quantity, is adequately represented by its physical correlate, fundamental frequency, for harmonic sounds such as singing (Vurma and Ross, 2006).

We relate fundamental frequency, $f_0$, to musical pitch, $p$, as follows:

$$p = 69 + 12\log_2\frac{f_0}{440}. \quad (1)$$

This scale is chosen such that a difference of one corresponds to one semitone; for integer pitches the representation coincides with the MIDI pitch scale, with reference pitch A4 tuned to 440 Hz ($p = 69$). As pitch differences are generally small, we often use the unit *cent*, equal to a hundredth of a semitone, in ET.

For example, middle C (60 on the MIDI pitch scale) has a frequency of 261.63 Hz. A note measured at 257 Hz has a

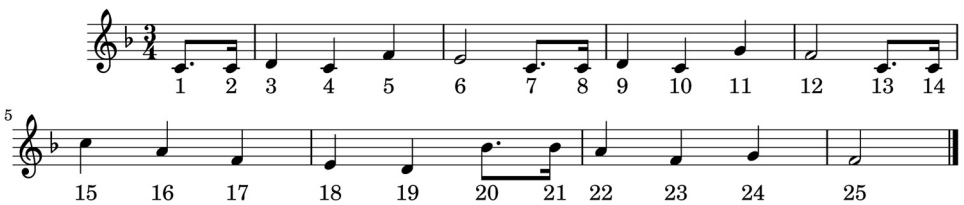



FIG. 1. "Happy Birthday" in F-major.

pitch of 59.69, and thus an intonation difference to middle C of −0.31 semitones (or −31 cents).

We use the word *nominal* to refer to the ideal intervals or pitches with respect to a reference in ET. For instance, if we consider an upward interval of a perfect fifth, then its nominal size is 7 semitones. This allows us to contrast this with the size of an observed interval, which, in general, differs from its nominal size.

The recorded songs were analyzed using a semi-automatic pitch tracking process. The second author (K.F.) annotated onsets and offsets of note events by visually identifying the stable part in the estimated pitch track using Sonic Visualizer 2.0 (Cannam *et al.*, 2010) and subsequent auditory verification. Automatically calculated onsets and offsets were adjusted manually, and the resulting annotations were fed into customized pitch tracking software (Mauch and Dixon, 2014), which is based on the YIN algorithm (de Cheveigné and Kawahara, 2002). The resulting note tracks were then analyzed using R (Team R Development Core, 2008). In order to obtain note-wise pitch estimates, we take the median pitch estimate over the annotated duration of the note, as illustrated in Fig. 2. A total of 4789 notes in 72 recordings were collected this way.

To test the reliability of the note timing annotations, 12 randomly selected blocks (of 3 runs) were also annotated manually by the other two authors and submitted to the note tracking algorithm. A comparison of onset and offset annotations reveals that these coders chose voiced/unvoiced boundaries and included note transitions, while K.F. consistently placed onsets later and offsets earlier in the sound event, capturing only the stable pitch portion of the note. A comparison of the different resulting pitch tracks showed that the median statistic is robust to such varying interpretations of note onsets and offsets, and no significant differences for the note pitch estimates were found. The average difference of the two other coders to the first coder was less than 0.2 cents, and only 1.4% of $F_0$-differences were larger than 5 cents.

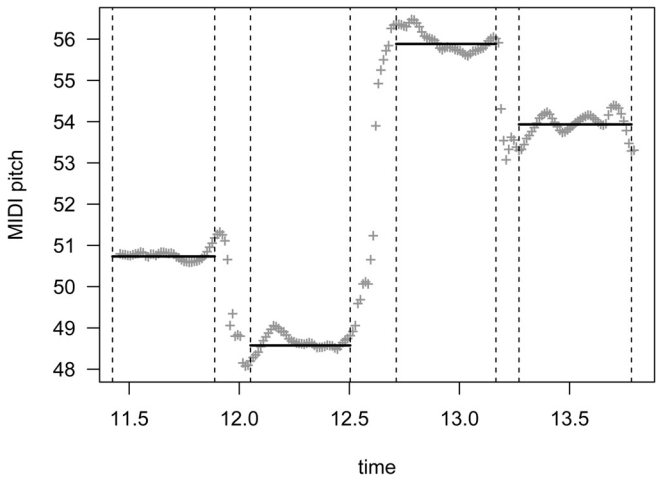



FIG. 2. Example pitch track (gray crosses) and note-wise pitch estimates (horizontal bars) calculated as medians between annotated note boundaries (vertical dashed lines).

## IV. METRICS OF ACCURACY AND DRIFT

In this section, we introduce how we measure intonation (in terms of interval and pitch error), singer-wise performance, and drift. We start by defining interval and pitch errors for individual notes and illustrate these using some examples from our data. Then, we introduce measures of intonation accuracy and drift based on the error definitions.

### A. Interval error

The distance between two pitches is referred to in musical terms as an interval, corresponding, in physical terms, to the ratio of the constituent fundamental frequencies. For the sake of this paper, we express the interval leading to the $i$th pitch, $p_i$ [see Eq. (1)], as the signed distance, $\Delta p_i = p_i - p_{i-1}$, in semitones between the $i$th and the preceding note. The interval error of the observed interval, $\Delta p_i$, can then be written as

$$e_i^{\text{int}} = \Delta p_i - \Delta p_i^0, \tag{2}$$

where $\Delta p_i^0$ is the nominal interval in semitones using ET. Figure 3(a) shows a box plot of interval error by nominal interval. A first observation is that the two largest upward intervals of 8 semitones (*minor sixth*) and 12 semitones (*octave*) are significantly flat, i.e., smaller than expected [one sample $t$ test: $t(186) = -6.96$, $t(183) = -9.09$, both $p < 0.0001$]. This phenomenon is called compression and is well known in the literature (Pfordresher *et al.*, 2010).

The *prime* interval, a repetition of the same pitch (0 semitone nominal interval), is systematically sharp, i.e., sung too high [one sample $t$ test: $t(753) = 17.96$, $p < 0.0001$] by ∼0.29 semitones. The fact that all prime intervals occur between the first and second note of each phrase [see Fig. 3(b)] suggests two possible explanations. Either the *first* note is sung flat as the vocal cords re-adjust from low tension in the rest between phrases to the higher tension required to sing the intended pitch, or the *second* note is sharp in preparation for an upward interval occurring after the note. This second possibility cannot explain the sharpness of note 21, which is followed by a downward interval, but we will obtain further insights by considering pitch error.

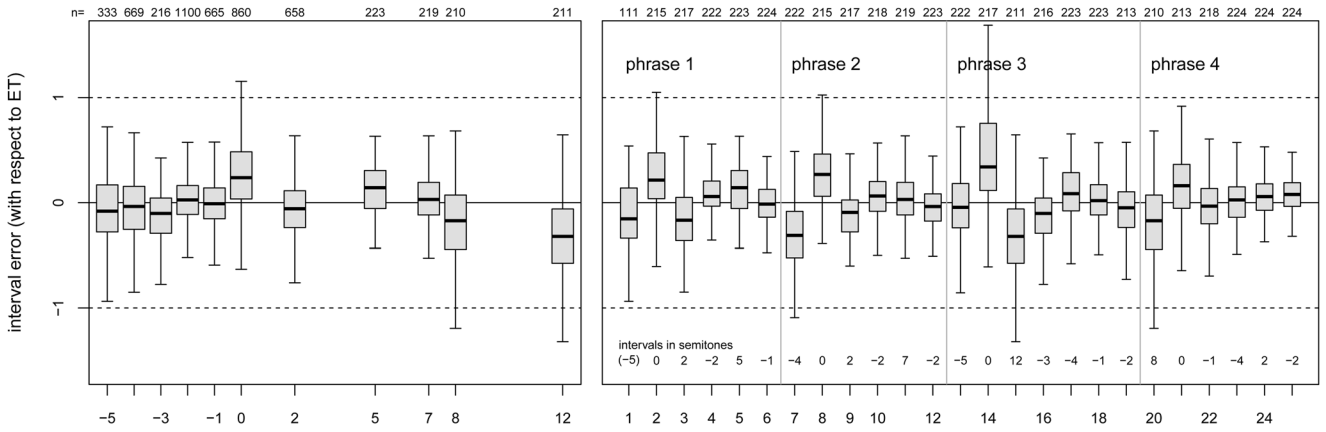
### B. Pitch error

Defining pitch error is not as straightforward as defining interval error because in our unaccompanied singing data we have no external reference pitch against which intonation could be measured. Instead, the tuning emerges as singers sing and may change over the course of the song. As a result, no single best way of defining intonation is possible.

In order to obtain a reference, we will use a linear fit to the local tonic estimate, as explained below. For the measured pitch, $p_i$, of the $i$th note, we can find an estimate,

$$t_i = p_i - s_i, \tag{3}$$

of the implied tonic pitch by subtracting from $p_i$ the nominal pitch, $s_i$, relative to the estimated tonic. These nominal pitches for "Happy Birthday" are given in Fig. 5(b). For example, if the first note in a run is sung at $p_1 = 50.45$ [see




(a) by nominal interval (in semitones) (b) by note

FIG. 3. Box-and-whisker plot of interval errors in semitones for all recordings of all 24 singers. Values are relative to the score using ET. The numbers at the top of the figure indicate sample sizes.

Eq. (1)], then the implied tonic is $t_1 = 50.45 - (-5) = 55.45$ because the first note is 5 semitones below the tonic. This is shown in Fig. 4, which also illustrates the next steps. For every run (a third of the performance), we use linear regression to fit a line to the 25 values $t_i$ with note number, $i$, as independent variable, obtaining fitted values, $t'_i$. (Linear regression was chosen as the simplest approach allowing for tonic changes.) We define the note error, $e_i$, as the difference between the implied tonic and the fitted tonic,

$$e_i = t_i - t'_i. \tag{4}$$

The individual errors are represented by the stems between the linear fit and the filled markers in Fig. 4.

With the ability to measure the pitch error, we can now investigate the relative effects of phrase beginnings and note jump preparation, as hypothesized in Sec. IV A. A linear model predicting pitch error by the independent variables *is-beginning-of-phrase* and *interval-to-next-note* shows

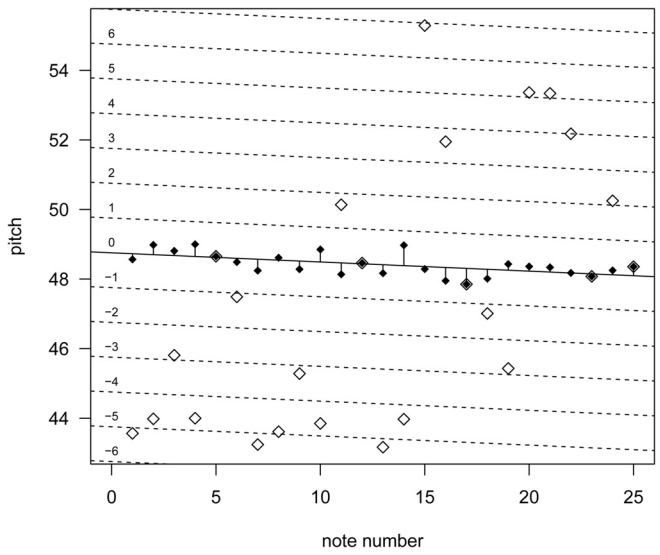



FIG. 4. Example of pitch error estimation, showing pitch measurements, $p_i$ (empty bullets), and local tonic estimates, $t_i$ (filled), using a linear fit. The stems represent the pitch error, $e_i$.

that both correlate significantly [$F(4667) = 254.94$, both $p < 0.0001$] with interval error. Hence, neither hypothesis can be rejected; it is likely that both influence intonation. Being at the beginning of a phrase "makes" notes about 21 cents flat. Each signed semitone in the following interval leads to a sharpening of 1.3 cents (upward octave example: $12 \times 1.3 = 15.6$ cents). Together, the two variables account for 9.8% of the variance (as measured by $R^2$).

While using other reference temperaments would be possible, they do not provide substantially differing errors, which is in line with previous results by Devaney *et al.* (2011). In fact, in terms of mean absolute pitch error (see Sec. IV C), ET is a significantly better hypothesis than just intonation [$t(4774) = -14.1927$, $p < 0.0001$], but the actual difference is very small (1.3 cents). Last, note that interval and pitch errors indicate deviation from the mathematically defined ET grid, not an esthetic judgment.

### C. Metrics of singing accuracy and precision

In order to assess singing accuracy, we use two metrics: mean absolute pitch error (MAPE), defined as

$$\text{MAPE} = \frac{1}{M}\sum_{i=1}^{M} |e_i|, \tag{5}$$

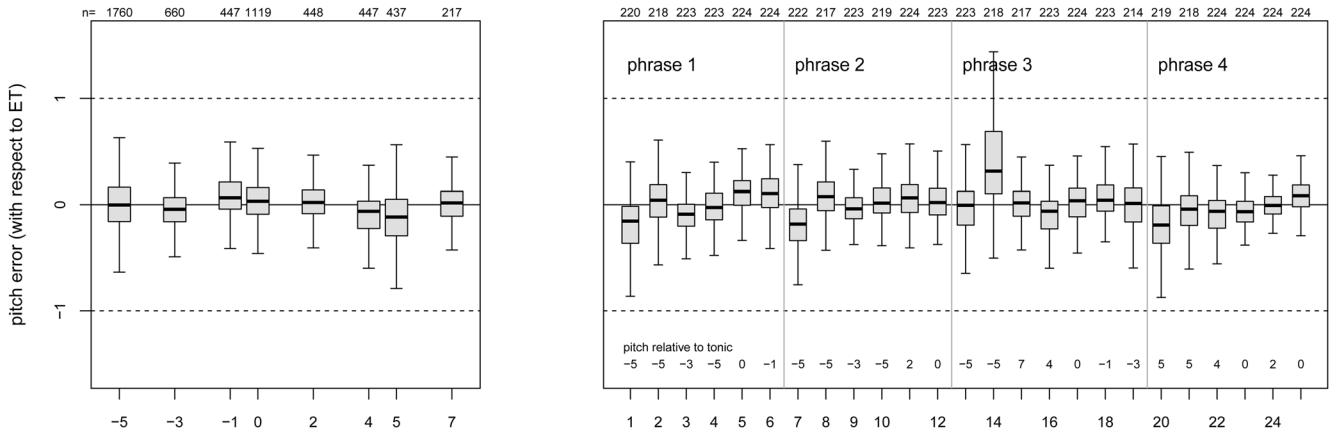
and mean absolute interval error (MAIE), defined as

$$\text{MAIE} = \frac{1}{M-1}\sum_{i=2}^{M} |e_i^{\text{int}}|. \tag{6}$$

Both metrics are always non-negative. MAIE does not reflect any tendency to sing larger or smaller intervals, but it is, in our view, a natural way to indicate how closely intervals match their target (and is equivalent to *interval deviation*; Dalla Bella *et al.*, 2007).

### D. Metrics of pitch drift

Each of our recordings has a first and third run of "Happy Birthday," each consisting of 25 notes. We estimate




(a) by pitch from tonic (in semitones)

(b) by note

FIG. 5. Box-and-whisker plots of pitch errors in semitones for all 24 singers with respect to linear prediction (run-wise). The numbers at the top of the figure indicate sample sizes.

drift based on pitch differences between corresponding notes in these two runs of the song. Hence, for a particular recording, we define pitch drift, $D$, as the mean difference,

$$D = \frac{1}{25}\sum_{i=1}^{25} p_{i+50} - p_i. \tag{7}$$

The drift metric, $D$, conveys information about the magnitude and direction of drift. In order to consider only the magnitude, we use the metric *absolute drift*, i.e., $|D|$, which is equivalent to *pitch stability* (Dalla Bella *et al.*, 2007; see also Flowers and Dunne-Sousa, 1990, p. 105).

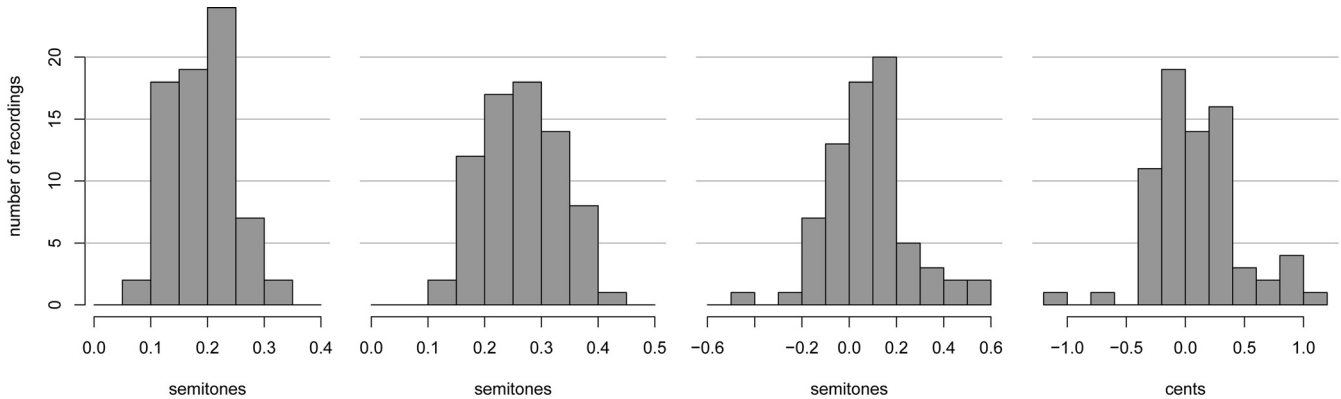
In the more general case, without repeated sequences, drift can be estimated as the slope of a linear model predicting the local tonic estimates, $t_i$, with the note numbers $1, \ldots, 75$ as the covariate. We have already used the same technique to calculate pitch error (Sec. IV B). As we will see in Sec. V, this *linear drift*, denoted $D_L$, is very highly correlated with $D$, so for most of our analyses, we will use only $D$ and $|D|$. From the model used to determine $D_L$ for a particular recording, we also calculate the associated $p$-value, which is an indicator of the significance of the drift effect.

## V. RESULTS

The metrics summarizing accuracy and drift defined in Sec. IV allow us to analyze recordings and assess the correlations with test condition (*Normal*, *Masked*, *Imagined*) and participant factors, such as choir experience. In order to prepare for the correlation analyses, we first present the distributions of recording-wise summary statistics themselves.

### A. Distributions of accuracy and drift

We calculated the MAPE (see Sec. IV C) for each of the 72 recordings. Figure 6(a) provides a histogram of the distribution of MAPE, showing that the average error magnitude is <0.5 semitones for all recordings, with most recordings having a MAPE of ∼0.2 semitones [mean: 0.189; median: 0.187; standard deviation (std. dev.): 0.051]. While this result shows that the singing in most recordings was systematically compatible with ET, it is also clear that 0.2 semitones (20 cents) is slightly larger than the just noticeable difference, which for typical singing frequencies up to 800 Hz is usually <1%, i.e., <17 cents (Henning, 1966). The distribution of MAIE [Fig. 6(b)] is similar, with slightly larger




(a) mean abs. pitch error

(b) mean abs. interval error

(c) drift $D$

(d) linear drift $D_L$

FIG. 6. Distributions of singing accuracy metrics over all conditions and participants.

magnitudes of ∼26 cents (mean: 0.263; median: 0.267; std. dev.: 0.069). Turning to Table II, we observe that MAPE and MAIE are indeed correlated, almost deterministically, across recordings (Spearman rank correlation of 0.93). What is remarkable is that neither significantly correlates with drift nor absolute drift. This suggests that the capability of remaining in a key does not depend on the ability to sing individual notes accurately. This conclusion is valid only if we can show that the drifts we observed are unlikely to stem from measurement error. The question is, hence, whether the drifts we do observe are statistically significant.

First, we consider the distribution of drift over recordings. A histogram of drift, $D$, is shown in Fig. 6(c) (in semitones, mean: 0.074; median: 0.069; std. dev.: 0.169) and linear drift, $D_L$, in Fig. 6(d) (in cents, mean: 0.097; median: 0.096; std. dev.: 0.371). The absolute intonation drift, $|D|$ (in semitones, mean: 0.138; median: 0.111; std. dev.: 0.122), has a mean of only 0.138, which is smaller than the mean MAPE (0.187). That is, in our sample, the expected drift magnitude over 50 notes is smaller than the expected absolute error per note.

In order to test whether the drifts are a real effect rather than measurement noise, we fit a recording-wise linear regression model to the implied tonic measurements, $t_i$, as described in Sec. IV C. For each recording, we obtain the $p$-value of the slope, with low values indicating strong evidence for the existence of significant drift. Figure 7 plots these $p$-values against linear drift, $D_L$. Of the 72 recordings, 16 (22%) have a $p$-value below the line of confidence level, 0.01; that is, they show significant drift. (Relaxing the confidence level to 0.05, significant drift occurs in 27 recordings, or 38%.) We conclude that drift is indeed a real effect. Hence, the lack of correlation between our measures of drift, on the one hand, and MAIE and MAPE, on the other, is a non-trivial finding.

In our dataset, the vast majority of recordings with significant drift actually drift upward. This is surprising especially because many choirs suffer from the opposite phenomenon, as discussed in Sec. I, but in line with some findings on solo folk singing (Müller *et al.*, 2010).

In summary, despite significant drift, drift effects are unrelated to the magnitude of pitch error and interval error. This is all the more surprising given that the magnitudes of MAPE and MAIE are so widely spread. For example, recordings with MAPE values as disparate as 0.1 semitones and 0.3 semitones can show very similar drift magnitudes near to zero. The relative independence of drift and local error is further emphasized by the fact that all have absolute values in the same order of magnitude, which is incompatible with an intonation model in which pitch errors propagate, as we will explore in Sec. VI. First, however, we investigate correlations with the singers' self-assessment and experimental conditions.

TABLE II. Spearman rank correlations of survey metadata (singing ability, singing experience, musical background, choir experience) and measures of accuracy and drift. Significant correlations ($p < 0.01$) are shown in bold.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| sg.abl | **0.40** | **0.31** | **0.54** | **−0.45** | **−0.46** | 0.11 | −0.02 | 0.06 |
| | sg.exp | −0.07 | **0.42** | −0.16 | −0.27 | 0.20 | 0.05 | 0.11 |
| | | mus.bg | **0.34** | −0.16 | −0.24 | 0.10 | −0.02 | 0.05 |
| | | | ch.exp | **−0.37** | **−0.40** | 0.22 | 0.01 | 0.07 |
| | | | | MAIE | **0.93** | −0.19 | −0.01 | −0.06 |
| | | | | | MAPE | −0.19 | −0.01 | −0.04 |
| | | | | | | $D_L$ | **0.52** | **0.94** |
| | | | | | | | $\|D\|$ | **0.54** |
| | | | | | | | | $D$ |

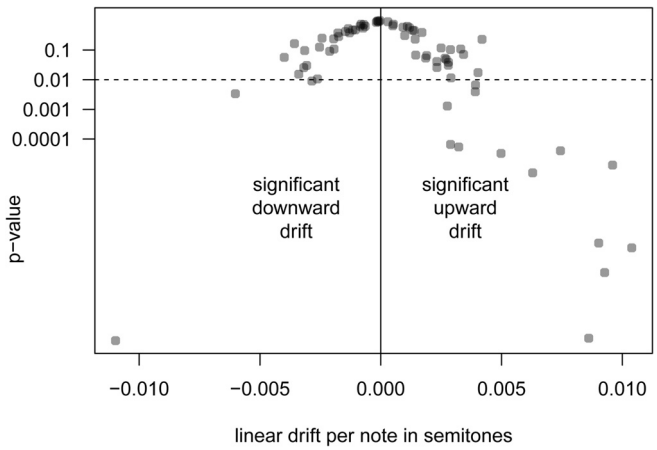



FIG. 7. Significance of drift, showing $p$-values (logarithmic scale) against $D_L$ for each recording. $p$-values below 0.01 are considered significant.

## B. Correlation with self-assessment

We investigated the relation between the quantitative intonation metrics and the singers' self-assessment, taken from a survey they filled in. Three self-reported metrics take values from 1 to 5: singing ability (poor to very high), singing experience (none to professional), choir experience (none to still active), and musical background (none to professional) takes values from 1 to 4. Table II shows the Spearman (i.e., rank) correlation values between all metrics, with significant correlations ($p < 0.01$) highlighted in bold print. We observe that most of the self-reported measures are inter-correlated, with the only exception of singing experience/musical background. In fact, the self-reported general level of musical background does not correlate with any of the quantitative measures either. Further study may reveal whether singing skills are indeed partially independent of general levels of musicality, as has been suggested before (Hutchins and Peretz, 2012).

However, two kinds of self-assessment ratings, singing ability and choir experience, do significantly correlate with our quantitative measures, MAPE and MAIE. All of the four combinations have absolute correlations ≥0.37. While the correlation of accurate singing and choir membership is expected, the singers' assessment of their singing ability, too, is in line with our measurements of intonation accuracy.

As we have mentioned in Sec. V A, we observed little correlation between the measures of accuracy, MAPE and MAIE, and measures of drift, $D$ and $|D|$. In fact, the only two metrics that correlate with drift, $D$, are those that are indeed directly related: linear drift, which is a different measure of the same phenomenon, and absolute drift, $|D|$, which correlates because most of the $D$ values are actually positive,

i.e., they coincide with $|D|$. Again, other than these direct connections, no other metrics correlate with either $D$ or $|D|$, in particular, none of the self-reported measures, including singing experience and choir experience.

### C. Effect of experimental conditions: *Normal*, *Masked*, *Imagined*

To see whether the three conditions (*Normal*, *Masked*, *Imagined*; see Sec. III) have an influence on our measures of accuracy and drift, an analysis of variance was conducted. Since all four accuracy and precision variables are not normally distributed (right-skewed), a set of non-parametric Kruskal-Wallis tests was performed, but no significant differences between conditions and runs were found [MAPE : $\chi^2(2) = 0.89$, $p = 0.64$; MAIE : $\chi^2(2) = 2.43$, $p = 0.30$; $D$ : $\chi^2(2) = 2.51$, $p = 0.28$; $|D|$ : $\chi^2(2) = 0.42$, $p = 0.81$]. Even the middle run in the *Masked* condition did not significantly deteriorate singing intonation, in contrast with some other findings (e.g., Mürbe *et al.*, 2002), but consistent with others who used low-level noise similar to that in our experiments (e.g., Pfordresher and Brown, 2007).

One observation during the experiments was that singers tend to sing louder in the *Masked* condition, compensating for the deprived auditory feedback (the so-called Lombard effect; Lombard, 1911), which is likely to have made the auditory feedback inhibition ineffective. The fact that the *Imagined* condition has little bearing on intonation is in line with perceptual experiments which found little difference in pitch acuity between listening and imagining conditions (Janata and Paroo, 2006). In summary, the conditions had no significant effect on the parameters we tested.

## VI. A MODEL FOR INTONATION STABILITY

In this section, we consider the question: how do singers stay in tune at all? While significant pitch drift was detected in many recordings, the tuning difference over three runs of "Happy Birthday" stayed remarkably small, despite large intonation errors on individual notes (see Sec. V A). It appears that even amateur singers possess a mechanism that prevents them from chaotically drifting out of tune. This stabilizing mechanism, we hypothesize, is mainly based on the retention of a pitch reference in short-term memory.

### A. Production with memory of a changing reference pitch

A simple pitch production model can be built on the assumption that the intonation of the $i$th note consists mainly of two components: a reference pitch, $r_i$, and the score information relative to that reference pitch. We choose to encode the melody notes in semitones relative to the tonic. (This is arbitrary; any other reference yields an equivalent model.) Assuming an additive Gaussian pitch error, $\varepsilon_i \sim N(0, \sigma_i)$, the pitch production process can then be written as

$$p_i = r_i + s_i + \varepsilon_i, \tag{8}$$

where $p_i$ is the pitch of the $i$th note, $r_i$ is the reference pitch, and $s_i$ is the fixed score information given relative to the tonic. The error, $\varepsilon_i$, models all additional noise, e.g., from physiological effects.

Our results on pitch drift (see Sec. V A) indicate that the singers' reference pitch changes over time. We assume that the memory of the pitch reference cannot be perturbed by future events and, hence, model $r_i$ as the causal process,

$$r_i = \mu r_{i-1} + (1-\mu)\,(p_{i-1} - s_{i-1}), \tag{9}$$

which depends on the previous reference pitch, $r_{i-1}$, and a point-estimate of the reference pitch $(p_{i-1} - s_{i-1})$, where $\mu \in [0,1]$ is a parameter relating to the memory of the previous reference pitch, $r_{i-1}$. Re-writing Eq. (9) as

$$r_i = r_{i-1} + (1-\mu)e_{i-1} \tag{10}$$

illustrates that the reference pitch is "pulled" in the direction of observed error, $e_{i-1} = (p_{i-1} - s_{i-1}) - r_{i-1}$. A similar model, based on updated tuning histograms, was proposed by Ryynänen (2004) to deal with the transcription of monophonic melodies in an engineering context.

Since no reference pitch is available before the first observation, Eq. (9) is not defined for $i = 1$, i.e., we have a cold start problem. We choose the first phrase (six notes) to initialize the smoothed reference pitch estimate, $r^* = \frac{1}{6}\sum_i t_i = \frac{1}{6}\sum_i (p_i - s_i)$. The first six notes in every recording are then excluded from any further analysis of this model, and the recursive update (9) is applied from $i = 7$. Figure 8 shows the local and smoothed reference pitches for an example recording under the *Normal* condition.

### B. Boundary models: No memory and absolute memory

The extreme cases, $\mu = 0$ and $\mu = 1$, generate models with no memory of the reference pitch (in the Markovian sense) and perfect memory of the reference pitch, respectively. If $\mu = 0$, only the previous note realization is used for reference, i.e., the reference pitch is simply $r_i = (p_{i-1} - s_{i-1})$, and hence

$$p_i = p_{i-1} + \underbrace{(s_i - s_{i-1})}_{\text{interval}} + \varepsilon_i.$$

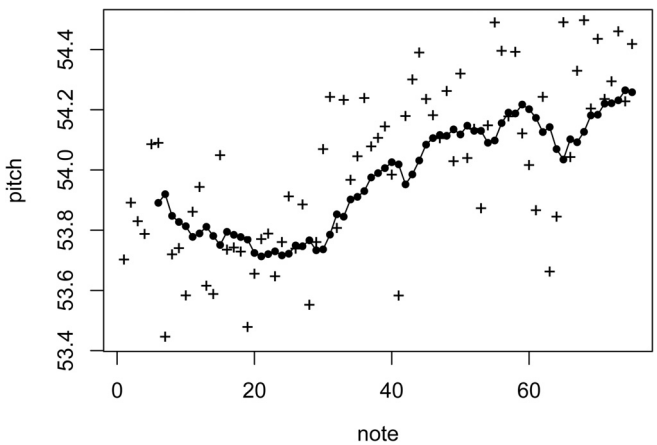



FIG. 8. Example of observed tonality estimates, $t_i$ (marked as +), and the estimated reference pitch, $r_i$ (filled bullets), with parameter $\mu = 0.85$.

That is, pitch production is based on the interval from the previous note realization. This also means that errors from the previous note are fully passed on. Mathematical formalization confirms that with an arbitrary starting pitch, $p_0$, the pitch variance, $\mathrm{Var}[p_i - p_0] = \sum_{j=1}^{i} \mathrm{Var}[\Delta p_j]$, is the sum of the interval error variances (assuming that intervals are independent). At the average observed interval variance of $\mathrm{Var}[\Delta p_i] = 0.147$, the expected variance of two notes spaced 50 notes apart is $50 \times \mathrm{Var}[\Delta p_i] = 7.36$. This corresponds to a standard deviation of 2.71 semitones, which is very clearly different from the 0.28 semitones standard deviation observed in our study (see Sec. V A).

The other extreme is $\mu = 1$, in which case the original reference pitch is perfectly maintained, and no information is passed on from one note to the next. In our case, the reference pitch remains $r^*$ throughout the piece. Given a fixed reference pitch, $r^*$, the constant reference pitch model predicts that the variance of the error, $t_i - r^*$, remains constant across a recording, which is another way of saying that no drift occurs. To test this prediction, we proceed as follows: we calculate the errors, $t_i - r^*$, with respect to the reference, $r^*$ (based on the first phrase, as in Sec. VI A), and estimate per-note variances across all recordings. We use a linear model with pitch error as covariate in order to subtract the linear effect of pitch error variances in individual notes. The resulting pitch-error-corrected residuals show a highly significant increase of variance with notes: note number explains 31.3% of the variance [$F(67) = 30.51$, $p < 0.0001$]. Over 75 notes, the standard deviation of residuals increases by 0.27 semitones. On these grounds, it is very unlikely that a constant reference pitch is used, and we have to reject the boundary model for $\mu = 1$.

Hence, both boundary models are at odds with our observations: one predicts extremely volatile drifts, the other—in its assumption of perfect reference pitch memory—predicts zero drift. The question is, then, whether a model with an intermediate value of $\mu \in (0, 1)$ will fit the data better.

### C. An intermediate memory parameter, $\mu$

Having rejected the boundary models for $\mu = 0$ and $\mu = 1$, we are interested in finding whether any intermediate $\mu$ provides a more adequate model. A good model should predict the observed individual note pitches with little error.

Since $r_i$ is meant to represent $t_i = (p_i - s_i)$ up to a note-wise error, as illustrated in Fig. 8, it seems plausible that, for some parameter, $\mu$, the prediction error can become small. We measure the model's MAPE (*model* MAPE) with respect to this reference. Figure 9 shows the error on a grid of $\mu$ values (equidistant with hop size 0.01). The best model is achieved for $\mu = 0.85$, leading to a *model* MAPE of 22 cents, with errors substantially higher toward the extremes of $\mu = 0$ (27 cents) and $\mu = 1$ (29 cents). While the figure shows that the linear model prediction is better (MAPE: 19 cents), only the memory model is psychologically plausible because it is causal, i.e., it does not depend on future events.

We also determined the $\mu$ values that minimize the error on individual recordings and averaged them by singer to obtain singer-wise $\mu$ values. Figure 10 shows a histogram of these singer-wise estimates, which range from $\mu = 0.62$ to $\mu = 0.98$ (mean: 0.832; median: 0.850; std. dev.: 0.105).

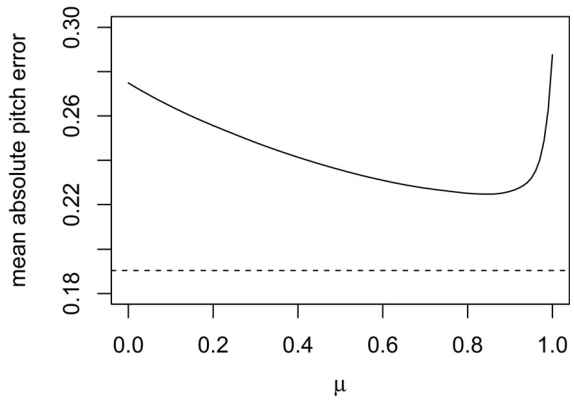



FIG. 9. Mean absolute error for models based on Eq. (9) for different values of the memory weight, $\mu$. An optimum is recognizable around $\mu = 0.85$. Dashed line: best linear prediction.

The model behavior in both pitch prediction and spread of drift suggests that a memory model such as the one defined by Eqs. (8) and (9) is reasonable for values around $\mu = 0.85$.

## VII. DISCUSSION AND FUTURE WORK

New knowledge of intonation drift may have implications for practitioners of singing, especially in choirs. Our findings in Sec. V A suggest that unaccompanied solo singing, without a harmonic context or interaction with other musicians, rarely results in pronounced intonation drift. The median of 11 cents drift observed is not only smaller than the mean absolute error per note, but also in the range of differences of concurrent pitches measured in choirs [10–15 cents, according to Ternström and Sundberg (1988)]. This adds further evidence to other causes for drift, such as the interaction between temperament and intonation in polyphonic singing (Devaney and Ellis, 2008; Howard, 2007).

In terms of individual singers, the intonation memory model presented in Sec. VI is particularly interesting because the parameter $\mu$ can reflect the capacity of a singer to stay in tune and that, unlike interval error, is not immediately obvious when a person starts to sing. With three recordings

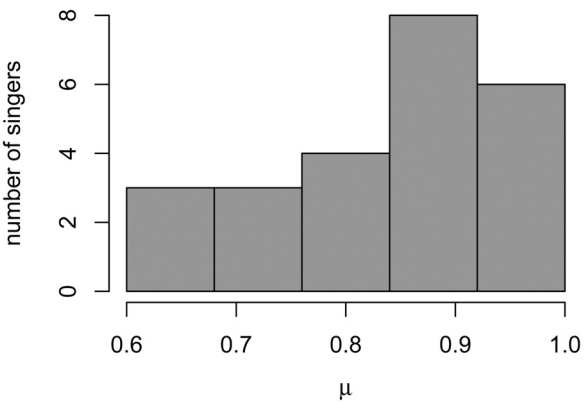



FIG. 10. Histogram of $\mu$ by singer.

per participant our data has allowed us to study some characteristics of individual singers, but more recordings of individual singers are necessary to refine our models and understanding of intonation memory. For example, our model is stationary, i.e., it predicts zero long term drift. A non-zero drift term might yield a more realistic model.

For this study, we chose to use "Happy Birthday" as our example tune, and while it is the most widely known song among non-professional singers, using only a single melody is an obvious limitation. More different melodies are needed to study intonation behavior in more detail and with more claim to generality.

While we found that, in our study, ET was as good a reference grid as just intonation, we hope that further experiments will enable us to infer more precisely the intonation intended by singers.

The analyses carried out in this paper all rely on individual notes as the fundamental musical unit. Future studies will include the temporal development of pitch within the duration of notes (e.g., glide, vibrato) and investigations on the effect of the duration itself.

## VIII. CONCLUSIONS

This paper has presented a study on intonation and intonation drift in unaccompanied solo singing. The main focus of the paper was the relations between drift (going out of tune), on the one hand, and measured pitch accuracy, different feedback conditions, and participants' self-assessment, on the other. Our main finding is that drift is common in solo singing. However, its extent is often small ($<0.2$ semitones over 50 notes) and not correlated to pitch accuracy, interval accuracy, or musical background. Most significant drifts, in our particular study, are upward drifts.

No significant difference was found between the three different singing conditions *Normal*, *Masked*, and *Imagined*, suggesting that, in our study, vocal strain and auditory feedback had little impact on the singers' capability of staying in tune.

Using our findings on solo intonation drift, we motivate a causal model of reference pitch memory with a single parameter, $\mu$, representing the memory strength. We show that values around $\mu = 0.85$ minimize the model MAPE.

The fact that significant drift occurs even in unaccompanied solo singing suggests that tuning changes in more complex situations, such as choir singing can partially be accounted for by drift. The small magnitude of drift observed in our study indicates that this is not inconsistent with earlier studies that highlight other causes.


Alldahl, P.-G. (**2006**). *Choral Intonation* (Gehrman, Stockholm, Sweden), p. 4.

Berkowska, M., and Dalla Bella, S. (**2009**). "Acquired and congenital disorders of sung performance: A review," Adv. Cogn. Psychol. **5**, 69–83.

Boersma, P. (**2002**). "Praat, a system for doing phonetics by computer," Glot Int. **5**, 341–345.

Brown, D. (**1991**). *Human Universals* (Temple University Press, Philadelphia, PA), pp. 1–160.

Cannam, C., Landone, C., and Sandler, M. (**2010**). "Sonic Visualiser: An open source application for viewing, analysing, and annotating music audio files," in *Proceedings of the ACM Multimedia 2010 International Conference* (Firenze, Italy), pp. 1467–1468.

Cano, E., Grollmisch, S., and Dittmar, C. (**2012**). "Songs2See: Towards a new generation of music performance games," in *9th International Symposium on Computer Music Modelling and Retrieval*, pp. 421–428.

Crowther, D. S. (**2003**). *Key Choral Concepts: Teaching Techniques and Tools to Help Your Choir Sound Great!* (Horizon, Springville, UT), pp. 81–85.

Dalla Bella, S., and Berkowska, M. (**2009**). "Singing proficiency in the majority," Ann. NY Acad. Sci. **1169**, 99–107.

Dalla Bella, S., Giguère, J.-F., and Peretz, I. (**2007**). "Singing proficiency in the general population," J. Acoust. Soc. Am. **121**, 1182–1189.

de Cheveigné, A., and Kawahara, H. (**2002**). "YIN, a fundamental frequency estimator for speech and music," J. Acoust. Soc. Am. **111**, 1917–1930.

Devaney, J., and Ellis, D. P. (**2008**). "An empirical approach to studying intonation tendencies in polyphonic vocal performances," J. Interdiscip. Music Stud. **2**, 141–156.

Devaney, J., Mandel, M., and Fujinaga, I. (**2012**). "A study of intonation in three-part singing using the automatic music performance analysis and comparison toolkit (AMPACT)," in *13th International Society of Music Information Retrieval Conference*, pp. 511–516.

Devaney, J., Wild, J., and Fujinaga, I. (**2011**). "Intonation in solo vocal performance: A study of semitone and whole tone tuning in undergraduate and professional sopranos," in *International Symposium on Performance Science*, pp. 219–224.

Filzmoser, P., Garrett, R., and Reimann, C. (**2005**). "Multivariate outlier detection in exploration geochemistry," Comput. Geosci. **13**, 579–587.

Flowers, P. J., and Dunne-Sousa, D. (**1990**). "Pitch-pattern accuracy, tonality, and vocal range in preschool children's singing," J. Res. Music Educ. **38**, 102–114.

Ganschow, C. M. (**2013**). "Secondary school choral conductors' self-reported beliefs and behaviors related to fundamental choral elements and rehearsal approaches," J. Music Teach. Educ. **20**, 1–10.

Henning, G. B. (**1966**). "Frequency discrimination of random-amplitude tones," J. Acoust. Soc. Am. **39**, 336–339.

Howard, D. M. (**2007**). "Intonation drift in a capella soprano, alto, tenor, bass quartet singing with key modulation," J. Voice **21**, 300–315.

Hutchins, S. M., and Peretz, I. (**2012**). "A frog in your throat or in your ear? Searching for the causes of poor singing," J. Exp. Psychol. Gen. **141**, 76–97.

Janata, P., and Paroo, K. (**2006**). "Acuity of auditory images in pitch and time," Percept. Psychophys. **68**, 829–844.

Kennedy, M. (**1980**). *The Concise Oxford Dictionary of Music* (Oxford University Press, Oxford, UK), p. 319.

Kleber, B., Zeitouni, A. G., Friberg, A., and Zatorre, R. J. (**2013**). "Experience-dependent modulation of feedback integration during singing: Role of the right anterior insula," J. Neurosci. **33**, 6070–6080.

Lombard, E. (**1911**). "Le signe de l'élévation de la voix" ("The sign of the elevation of the voice"), Ann. Mal. Oreil. Larynx **2**, 101–109.

Markel, J. (**1972**). "The SIFT algorithm for fundamental frequency estimation," IEEE Trans. Audio Electroacoust. **20**, 367–377.

Mauch, M., and Dixon, S. (**2014**). "PYIN: A fundamental frequency estimator using probabilistic threshold distributions," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2014)*, pp. 659–663.

Mithen, S. J. (**2007**). *The Singing Neanderthal: A Search for the Origins of Art, Religion, and Science* (Harvard University Press, Cambridge, MA), Chap. 16, pp. 246–265.

Molina, E. (**2012**). "Automatic scoring of singing voice based on melodic similarity measures," Master's thesis, Universitat Pompeu Fabra, Barcelona, Spain, pp. 9–14.

Müller, M., Grosche, P., and Wiering, F. (**2010**). "Automated analysis of performance variations in folk song recordings," in *Proceedings of the International Conference on Multimedia Information Retrieval*, pp. 247–256.

Mürbe, D., Pabst, F., Hofmann, G., and Sundberg, J. (**2002**). "Significance of auditory and kinesthetic feedback to singers pitch control," J. Voice **16**, 44–51.

Pfordresher, P. Q., and Brown, S. (**2007**). "Poor-pitch singing in the absence of 'tone deaf ness,' " Music Percept. **25**, 95–115.

Pfordresher, P. Q., Brown, S., Meier, K. M., Belyk, M., and Liotti, M. (**2010**). "Imprecise singing is widespread," J. Acoust. Soc. Am. **128**, 2182–2190.

Pinker, S. (**2002**). *The Blank Slate* (Viking, New York), p. 437.

Ryynänen, M. P. (**2004**). "Probabilistic modelling of note events in the transcription of monophonic melodies," Master's thesis, Tampere University of Technology, Finland, pp. 27–30.

Schroeder, M. R. (**1968**). "Period histogram and product spectrum: New methods for fundamental-frequency measurement," J. Acoust. Soc. Am. **43**, 829–834.
Seashore, C. E. (**1914**). "The tonoscope," Psychol. Monogr.s **16**, 1–12.
Seashore, C. E. (**1967**). *Psychology of Music* (Dover, New York), pp. 254–272.
Seaton, R., Pim, D., and Sharp, D. (**2013**). "Pitch drift in a cappella choral singing," Proc. Inst. Acoust. Ann. Spring Conf. **35**, 358–364.
Swannell, J. (**1992**). *The Oxford Modern English Dictionary* (Oxford University Press, New York), p. 560.
Team R Development Core (**2008**). *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria). Available at http://www.R-project.org (Last viewed September 9, 2013).
Terasawa, H. (**2004**). "Pitch drift in choral music," Music 221A final paper, Center for Computer Research in Music and Acoustics, at Stanford University, CA. Available at https://ccrma.stanford.edu/~hiroko/pitchdrift/paper221A.pdf (Last viewed 5 June 2014).
Ternström, S., and Sundberg, J. (**1988**). "Intonation precision of choir singers," J. Acoust. Soc. Am. **84**, 59–69.
Vurma, A., and Ross, J. (**2006**). "Production and perception of musical intervals," Music Percept. **23**, 331–344.
Welch, G. F. (**1985**). "A schema theory of how children learn to sing in tune" Psychol. Music **13**, 3–18.