

**Intonation in Unaccompanied Singing:
Accuracy, Drift and a Model of Intonation Memory**

Matthias Mauch,^{a)} Klaus Frieler,^{b)} and Simon Dixon^{c)}

*Centre for Digital Music,
Queen Mary University of London*

(Dated: September 2, 2013)

Abstract

This paper presents a study on intonation and intonation drift in unaccompanied human singing and proposes a simple intonation memory model that accounts for many of the effects observed. Singing experiments were conducted with 24 singers of varying ability under 3 conditions. Several summary measures of pitch and interval accuracy are revised and metrics for intonation drift and pitch stability are proposed. A median absolute drift of 11 cents was observed, which was significant in 22% of recordings. Drift magnitude did not correlate with other measures of singing accuracy, singing experience or with 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 behaviour observed. The proposed causal model provides a better fit.

PACS numbers: 43.75.Bc, 43.75.Rs, 43.70.Fq

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: singing is one of the human universals (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 is aimed at providing some insights into a parameter in singing that is crucial to many singing styles but has so far received little academic attention: intonation.

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:

$$\text{pitch difference} = 12 \log_2 \frac{f_0}{f_{ref}}, \quad (1)$$

where f_0 is the measured fundamental frequency and f_{ref} is the fundamental frequency of the reference (“correct”) pitch. (Note that we assume pitch, a perceptual quantity, is adequately represented by its physical correlate, fundamental frequency, for harmonic sounds such as singing (Vurma and Ross, 2006).) As pitch differences are generally small, we often use the unit *cent*, equal to a hundredth of a semitone in equal temperament. Semitones and cents are commonly used in research on pitch (e.g. Pfordresher and Mantell, 2012) because they are easier to interpret in musical terms than raw frequency differences, and the logarithmic frequency scale corresponds more closely to pitch perception than a linear scale does.

^{a)}Electronic address: `matthias.mauch@eecs.qmul.ac.uk`

^{b)}Electronic address: `klaus.frieler@hfm-weimar.de`; Also at Musikwissenschaftliches Institut, HfM Franz Liszt Weimar.

^{c)}Electronic address: `simon.dixon@eecs.qmul.ac.uk`

For this paper, we use as our reference tuning system equal temperament with A4 tuned to 440Hz. Then, adopting the MIDI pitch scale, which assigns an integer for each note of the chromatic scale, with middle C assigned 60 and A4, 9 semitones higher, being assigned 69, we define semitone pitch p as

$$p = 69 + 12 \log_2 \frac{f_0}{440}. \quad (2)$$

Thus we can map any fundamental frequency to a pitch in semitones. In this system, a nominal G has a frequency of almost exactly 392Hz, and hence a note G sung at 401Hz would have an intonation difference of $1200 \log_2 \frac{401}{392} \approx 39$ cents. We will see in Section IV that for our unaccompanied singing data, neither the tuning system nor the reference pitch is known, but that for the purposes of our study we can assume equal temperament at an estimated tuning frequency, without substantially affecting our results.

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 conducted with a group of singers of different abilities. Section IV defines and illustrates several metrics of singing accuracy and drift. In the results section (V) we show what intrinsic and external factors influence our measurements. The following section (VI) introduces a simple model of tonal memory which is able to account for the intonation stability and drift we observed. Section VII provides a discussion of achievements and future work, and a summary of our conclusions is found in section 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 laborious signal processing 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 feasible 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. Berkowska and Dalla Bella, 2009; Dalla Bella and Berkowska, 2009; Dalla Bella *et al.*, 2007; Pfordresher *et al.*, 2010).

Vurma and Ross (2006) investigated professional singers’ ability to sing three intervals (minor second, tritone and perfect fifth), 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 judgements were not significantly different from other expert listeners. Judgements 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 call 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 characterise 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) distinguish the accuracy (ability to reproduce 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 greater than one semitone), but only 13% of singers were inaccurate (absolute value of average error greater than 1 semitone). It was also observed that errors were greater for the imitation task than for a recall task. In our study, we use their definition of poor singers in order to exclude some singers (Section III), and we discuss their metrics for local intonation performance in Appendix A.

Most existing research on intonation is concerned with a fixed tuning system, but some authors have also studied temporal changes in reference pitch, which we call intonation drift. 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) report only negligible effects on the original Benedetti composition, while Howard (2007) reports drifts roughly in line with predictions on specially composed new pieces.

Far from being a purely theoretical concept, intonation drift is a daily practical concern of choir conductors, as they rehearse and perform pieces that are not designed to produce intonation drift (assuming benign composers). Conductors and singers have put forward various hypotheses in discussion forums (e.g. Barbershop Tuning Discussion, 2012) and in music tutorials (Crowther, 2003), attempting to explain the phenomenon of drift, but these explanations all still await empirical testing.

This study investigates intonation drift in unaccompanied solo singing, without further constraints such as consonance with other singers. In contrast with some of the work described above, the singers in our experiment refer only to their own memory to stay in tune, i.e. the pitch reference is solely internal. It is well-known among singing teachers that in similar situations (such as unaccompanied sight-singing practice) even accurate singers tend to drift. Therefore, our work investigates both singing accuracy and intonation drift with a



FIG. 1: “Happy Birthday” in F-Major

particular emphasis on understanding how they are related.

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. Three of the subjects were unable to sing the requested melody and were excluded from the study. Also excluded were four further participants, who were detected as outliers and hence classified as “poor” singers, an established phenomenon (Pfordresher and Brown, 2007). The outlier classification was performed using multivariate outlier detection (Filzmoser *et al.*, 2005) on two singer-based metrics: mean absolute interval error (see Section 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 years (mean: 34.5). The gender ratio was imbalanced with 6 females and 18 males in the sample. The musical experience of participants was wide-spread. Fourteen singers considered themselves amateur musicians, 9 professionals or semi-professionals, and 1 reported no musical background. Thirteen participants reported “a lot” of singing experience, 9 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 8 had previous choir experience, and 5 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 call paradigm with a variety of subjects from two different countries, the choice fell on “Happy Birthday”, probably the single best-known and most wide-spread 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). Hence it poses some intonation challenges even for experienced singers. 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. 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 using Audacity 2.0 running on a Windows Laptop or a MacBook Pro. A conventional headset (Logitech USB Headset 390) functioned both as microphone and headphones, through which participants were provided with the click track and the 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 of about 70-80 dB SPL 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 to sing, and to resume singing at the start of the third rendition.

The *Masked* condition diminishes auditory feedback, making it harder for the participants to hear their own singing, while the *Imagined* condition removes both auditory and kinesthetic feedback, i.e. the participants can neither hear their singing nor feel singing-induced movements and the states of their vocal cords (unless they moved their vocal cords without singing).

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 two syllables at the same pitch; this extra note was disregarded in the analysis. One singer consistently missed note 19.

D. Analysis

The recorded songs were analysed using a semi-automatic pitch tracking process. The resulting note tracks were then analysed using R (2008). Onsets and offsets of the steady states of note events were annotated using Sonic Visualiser 2.0 (Cannam *et al.*, 2010) by the second author (**kf**). Automatically calculated onsets and offsets were adjusted manually, and the resulting annotations were fed into customised pitch tracking software (<http://code.soundsoftware.ac.uk/projects/yintony> [will be made available upon publication]), which is based on the YIN algorithm (de Cheveigné and Kawahara, 2002). In order to obtain note-wise pitch estimates we take the median pitch estimate over the annotated duration of the note, as illustrated in Figure 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 a different strategy

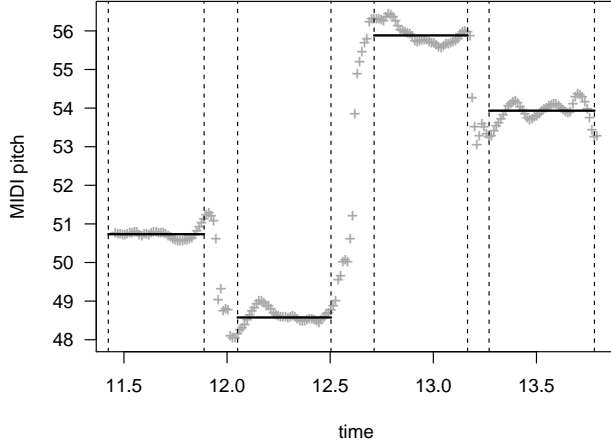


FIG. 2: Example pitch track (grey crosses) and note-wise pitch estimates (horizontal bars), calculated as medians between annotated note boundaries (vertical dashed lines).

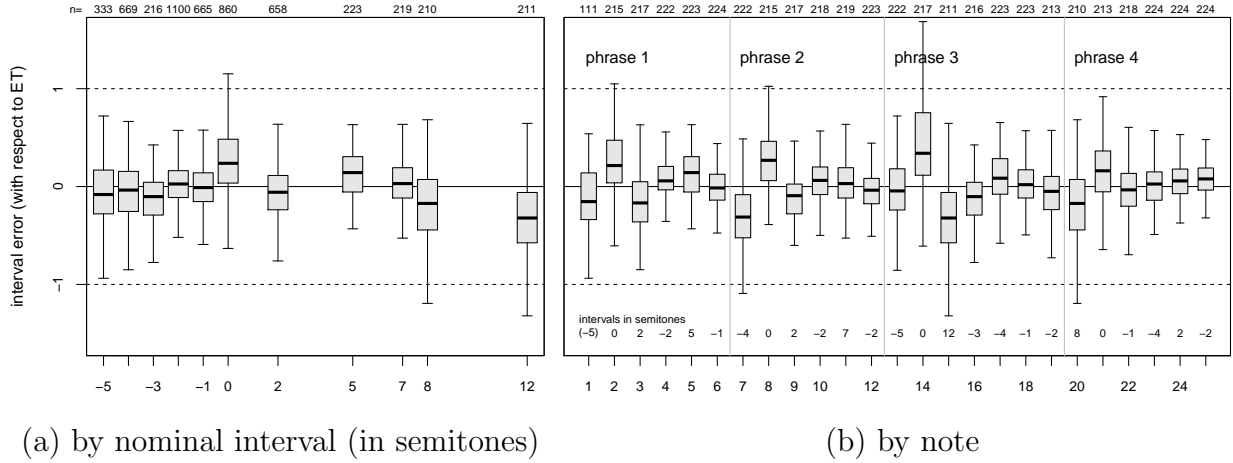
of coder **kf** compared to the other two: **kf** consistently placed onsets later and offsets earlier in the sound event, capturing only the steady state portion of the note. A comparison of the different resulting pitch tracks revealed no significant differences for the note pitch estimates. The average difference of the two other coders to the first coder was less than 0.2 cents, and only 1.4% of F0-differences were larger than 5 cents, showing that the median method of determining the pitch is robust against interpretations of note onsets and offsets.

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 measures 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



(a) by nominal interval (in semitones)

(b) by note

FIG. 3: Interval errors in semitones relative to the score, using equal temperament. The boxes indicate the 1st, 2nd (median) and 3rd quartiles, the whiskers extend to ‘the most extreme data point which is no more than 1.5 times the interquartile range’ (R software Team R Development Core (2008)). Outliers are omitted for clarity of display.

sake of this paper, we express the interval leading to the i^{th} pitch p_i (see Eq. (2)) as the signed distance $\Delta p_i = p_i - p_{i-1}$ in semitones between the current and the preceding note. The interval error of the observed interval Δp_i can then be written as

$$e_i^{\text{int}} = \Delta p_i - \Delta p_i^0, \quad (3)$$

where Δp_i^0 is the nominal interval in semitones using equal temperament (ET). Interval error is easy to define (unlike pitch error, see Section IV.B). Figure 3a 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).

A more puzzling case is the error at the nominal interval of zero semitones. In our data, this so-called *prime* interval, a repetition of the same pitch, is systematically sharp, i.e. sung too high (one sample t test: $t(753) = 17.96$, $p < 0.0001$) by approximately 0.29 semitones.

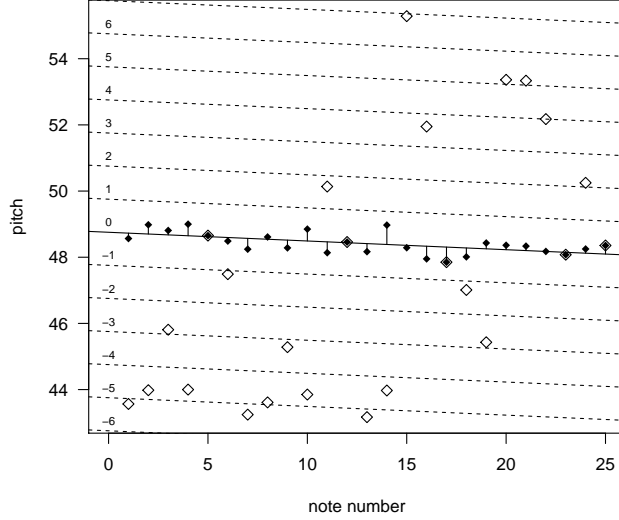


FIG. 4: Example of pitch error estimation, showing pitch measurements p_i (empty bullets) and local tonic estimates t_i (filled) using a linear fit. The stems represent the pitch error e_i .

Figure 3b suggests two possible explanations. Notice that all zero-semitone intervals occur between the first and second note of each phrase; they are the two notes that correspond each time to the lyric ‘*hap-py*’. A first hypothesis, then, is that the reason for the interval being sung sharp is that its *first* note is usually sung flat simply by virtue of being the starting note after the short pause between phrases, where the voice rests. A second hypothesis, that the *second* note is sharp in preparation for an upward interval occurring after the note, cannot explain the sharpness of note 21, which is followed by a downward interval. However, to test which hypothesis explains the data better, we need the concept of note intonation error, which is the topic of the next paragraphs.

B. Pitch error

Defining pitch error is not as straight-forward 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 pitch intonation is possible.

In order to obtain a reference we will use a linear fit to the local tonic estimate, as

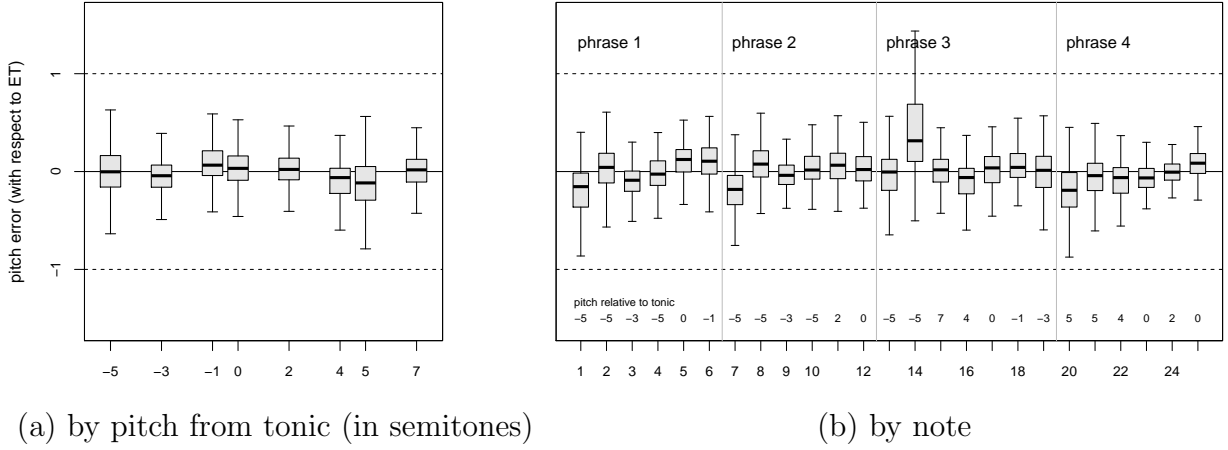


FIG. 5: Pitch errors with respect to linear prediction (run-wise).

explained below. For the measured pitch p_i of the i^{th} note we can find an estimate

$$t_i = p_i - s_i \quad (4)$$

of the implied tonic pitch by subtracting from p_i the nominal pitch s_i relative to the nominal tonic. These nominal pitches for “Happy Birthday” are given in Figure 5b. For example, if the first note in a run is sung at $p_1 = 50.45$ (see Eq. 2), 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 Figure 4, which also provides an intuitive illustration of the next steps: for every run we fit a line to the t_i , $i = 1, \dots, 25$ with note number i as independent variable, obtaining fitted values t'_i , $i = 1, \dots, 25$. We define the note error e_i as the difference between the implied tonic and the fitted tonic:

$$e_i = t_i - t'_i. \quad (5)$$

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

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

and *interval-to-next-note* shows that both correlate significantly ($F(4667) = 254.94$, both $p < 0.0001$) with interval error: beginning of phrase ‘makes’ notes about 21 cents flat. On average, each signed semitone in the following interval leads to a sharpening (or flattening, in case of downward intervals) of 1.3 cents, i.e. $12 \times 1.3 = 15.6$ cents in the case of the octave jump, $(-4) \times 1.3 = -5.2$ cents in the case of a downward major third. Together, they account for 9.8% of the variance (as measured by R^2). Hence, neither hypothesis can be rejected—it is likely that both influence intonation. The individual models explain less variance: beginning of phrase explains 3.8% ($F(4668) = 182.3$, $p < 0.0001$); following interval explains 6.2% ($F(4773) = 316.8$, $p < 0.0001$).

Note that neither interval nor pitch error can be used directly to judge the value or musical correctness of a sung note. Rather than a value judgement, ‘error’ indicates deviation from the mathematically defined equal temperament grid. 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 Section IV.C), equal temperament 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).

C. Metrics of singing accuracy and precision

Pfordresher *et al.* (2010) define four different metrics to summarise singing precision and accuracy in a recording or for a singer. However, in applying the measures to our data we encountered several problems arising from the definitions, both in terms of their intuitive understanding (the names are misleading) and power to express features of singing on our data (they obfuscated relevant information). Detailed explanations and definitions are given in Appendix A. In the following we therefore propose the use of alternative, more intuitive summary metrics.

A measure that combines pitch accuracy and precision is computed by averaging absolute

differences between rendition and target, which reflects intonation skill. Hence, we define the *mean absolute pitch error* (MAPE) as:

$$MAPE = \frac{1}{M} \sum_{i=1}^M |e_i|. \quad (6)$$

Similarly, our alternative measure to Pfordresher’s interval accuracy, the *mean absolute interval error* (MAIE) is defined as:

$$MAIE = \frac{1}{M-1} \sum_{i=2}^M |e_i^{\text{int}}|. \quad (7)$$

This measure is always non-negative, hence no tendency to sing larger or smaller intervals is reflected here, but it is in our view a more natural way to indicate how accurately intervals are sung.

D. Metrics of pitch drift

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

$$D = \frac{1}{25} \sum_{i=1}^{25} p_{i+50} - p_i. \quad (8)$$

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. $|D|$.

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 $1, \dots, 75$ as the covariate. We have already used the same technique to calculate pitch error (Section IV.B). As we will see in the following section, 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.

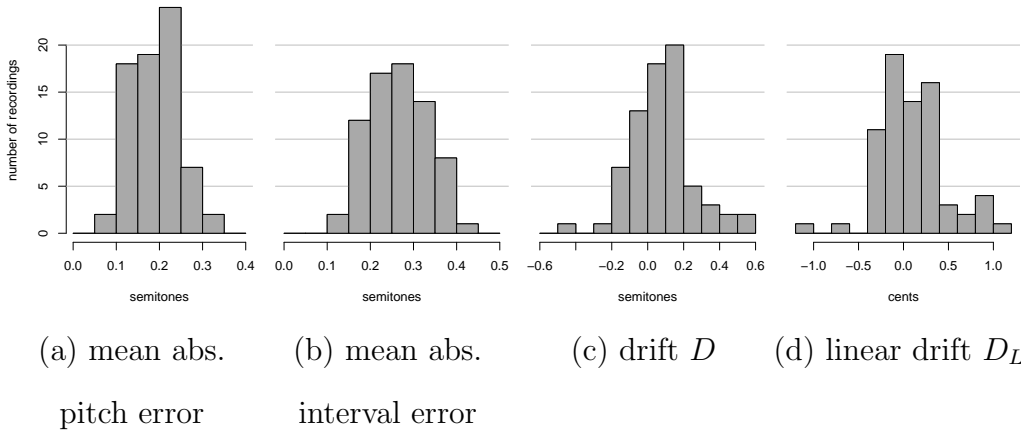


FIG. 6: Distributions of singing accuracy metrics over all conditions and participants.

V. RESULTS

The metrics summarising accuracy and drift defined in Section IV allow us to analyse 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 mean absolute pitch error (*MAPE*, see Section IV.C) for each of the 72 recordings. Figure 6a provides a histogram of the distribution of *MAPE*, showing that the average error magnitude is less than 0.5 semitones for all recordings, with most recordings having a *MAPE* of around 0.2 semitones (mean: 0.189; median: 0.187; std. dev.: 0.051). While this result shows that the singing in most recordings was systematically compatible with equal temperament, 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 800Hz is usually below 1%, i.e. below 17 cents (Henning, 1955). The distribution of *MAIE* is similar, with slightly larger magnitudes of around 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 or 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 Figure 6c (in semitones, mean: 0.074; median: 0.069; std. dev.: 0.169) and of linear drift D_L in Figure 6d (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 use the p value of the linear fit to the t_i values, as described in Section IV.C. Figure 7 plots the p value 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, 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.

A further, unexpected discovery is that—in our dataset—the vast majority of recordings with significant drift actually drift upwards. This is surprising especially because many choirs suffer from the opposite phenomenon (they tend to go flat).

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 emphasised by the fact that all have

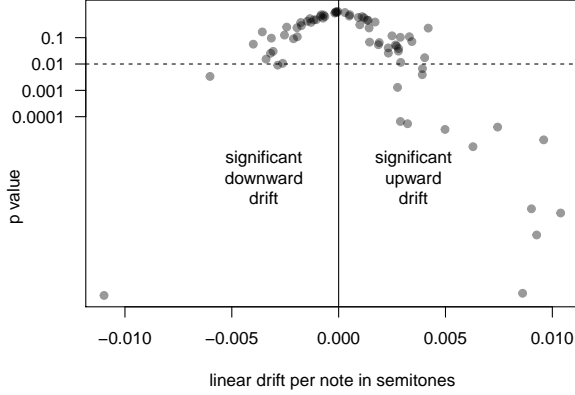


FIG. 7: Significance of drift, showing p values (logarithmic scale) against D_L for each recording. p values below 0.01 are considered significant.

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 Section VI. First, however, we investigate how different participant factors and singing conditions influence the results.

B. Participant factors

In order to determine whether knowing the singer constitutes a significant advantage to predicting $MAPE$, $MAIE$, D or $|D|$ in a recording we construct linear models predicting these metrics with singer as a (nominal) independent variable. The singer’s identity is indeed an excellent predictor for the $MAPE$ and $MAIE$ measures, explaining 85% ($F(23) = 12.19$, $p < 0.0001$) and 81% ($F(23) = 8.89$, $p < 0.0001$) of the variance between recordings. Predicting drift is much less effective with less variance explained and higher p values (in the case of D : 48%, $F(23) = 1.89$, $p = 0.03$; in the case of D_L : $F(23) = 2.75$, $p = 0.0015$), suggesting slightly lower predictive power. For absolute pitch drift $|D|$ prediction completely fails (only 22% of variance explained, $F(23) = 0.60$, $p = 0.91$). In short, though knowing the singer is likely to provide some information on the drift performance, the singers differ much more in terms of their ability to accurately sing single notes than in terms of characteristic

drift direction or magnitude. This result consolidates the qualitative difference between measures of accuracy and drift found in Section V.A.

We also 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) and 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 Section 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. Treatment factors

To see whether the three conditions (*Normal*, *Masked*, *Imagined*, see Section 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 in line 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.

In summary, the three conditions had little effect on the singers we tested. Neither their singing accuracy nor their tendency to drift were significantly affected.

D. Recapitulation

The various results from this section support the overarching impression that intonation drift is relatively independent of singers’ capability to sing individual notes accurately. About 22% of our recordings show a significant difference (drift) between the first and third run of “Happy Birthday”. The range of drifts, however, is small; for example, the mean absolute interval error is on the same order of magnitude as the drift over as many as 50 notes. Thus singers must possess a strong intonation memory which enables them to stay in tune. The next section proposes a model of intonation memory that is compatible with our findings.

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 Section V.A). It appears that even amateur singers possess a mechanism that prevents them from chaotically drifting out of tune.

This stabilising mechanism, we hypothesise, is mainly based on the short-term memory of a pitch reference. Before we introduce how we model this memory, we consider a basic model of pitch production.

A. Pitch Production under Constant Reference Pitch

A simple pitch production model can be built on the assumption that the intonation of a note consists mainly of two components: a reference pitch r , 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

$$p_i = r + s_i + \varepsilon_i, \tag{9}$$

where p_i is the pitch of the i^{th} note, r is the reference pitch and s_i is the fixed score information given relative to the reference pitch. The error ε_i models all additional noise, e.g. from physiological effects.

To illustrate the model, a baritone can sing comfortably in the pitch range around G3, so let us assume a reference pitch $r = 55.43$, corresponding to the tonic of “Happy Birthday”. The third note of “Happy Birthday” (‘hap-py *birth*-...’) is three semitones below the tonic, i.e. the score value is $s_3 = -3$. Then the desired sung note would be $r + s_3 = 55.43 - 3 = 52.43$. This process clearly captures important aspects of the pitch production process. However, its assumption of a static reference pitch would require the

singers to have perfect pitch memory, which, in general, is not the case.

B. Pitch Production Model with Imperfect Pitch Memory

The pitch drifts we observe in our data (see Section V.A) clearly indicate that singers do not retain a fixed reference pitch or tonality; rather, they slightly drift up and down while they sing, which indicates imperfect pitch memory. In order to capture imperfect pitch memory, we have to make adjustments to the model presented in Eq. (9) above.

Since the score notes s_i are fixed, we can extend the equation most naturally by modelling the tuning pitch r as a time-varying process r_i , i.e. the production equation now becomes

$$p_i = r_i + s_i + \varepsilon_i, \quad (10)$$

We assume that the process r_i is causal, i.e. it only depends on past events at times $j = 1, \dots, i-1$. In particular, a singer cannot predict the time-varying reference pitch from future local pitch deviations, so a linear model like the one used for the calculation of pitch error (see Section IV.B and Figure 4) with note numbers as covariates is not feasible. Instead, we assume that r_i is a causal smoothing process defined as the running mean

$$r_i = \mu r_{i-1} + (1 - \mu) (p_{i-1} - s_{i-1}) \quad (11)$$

of the memory reference 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 strength of memory. By calculating the running mean the influence of past notes decays geometrically. The recursive equation (11) is a simplistic model of a tuning memory process that pulls the reference pitch in the direction of the observed error $e_{i-1} = (p_{i-1} - s_{i-1}) - r_{i-1}$ at every step and can be re-written as

$$r_i = r_{i-1} + (1 - \mu)e_{i-1}. \quad (12)$$

A similar model, based on updated tuning histograms, was proposed by Ryyänen (2004) to deal with the transcription of monophonic melodies in an engineering context. Since no reference pitch is available before the first observation, Eq. (11) is not defined for $i = 1$,

i.e. we have a cold start problem. We choose the first phrase (six notes) to initialise the smoothed reference pitch estimate $r^* = \frac{1}{6} \sum t_i = \frac{1}{6} \sum (p_i - s_i)$. The first six notes in every recording are then excluded from any further analysis of this model, and the recursive update (11) is applied from $i = 7$. Figure 8 shows the local and smoothed reference pitches for an example recording under the *Normal* condition.

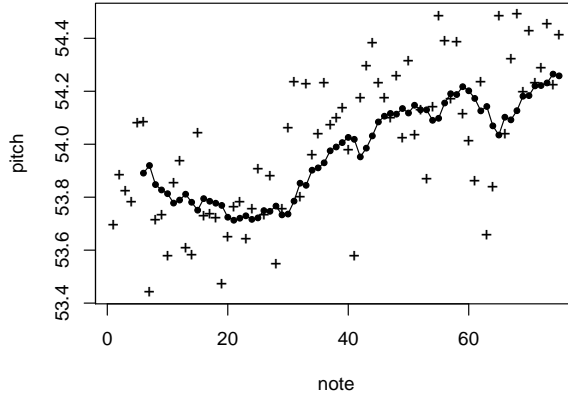


FIG. 8: Example of observed tonality estimates t_i (marked as +) and the estimated reference pitch r_i (filled bullets) with parameter $\mu = 0.85$.

C. Boundary models: no memory and absolute memory

The extreme cases $\mu = 0$ and $\mu = 1$ generate models with no memory and perfect memory, respectively. If $\mu = 0$, no memory is used to predict the current note except for the previous note realisation, i.e. the reference pitch is simply $r_i = (p_{i-1} - s_{i-1})$, and hence

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

That is, pitch production is based on the interval from the previous note realisation. This also means that errors from the previous note are fully passed on. Mathematical formalisation confirms that with an arbitrary starting pitch p_0 the pitch variance $\text{Var}[p_i - p_0] = \sum_{j=1}^i \text{Var}[\Delta p_j]$ is the sum of the interval error variances (assuming that

intervals are independent). At the average observed interval variance of $\text{Var}[\Delta p_i] = 0.147$ the expected variance of two notes spaced 50 notes apart is $50 \times \text{Var}[\Delta p_i] = 7.36$. 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 Section V.A).

The other extreme is $\mu = 1$, in which case *only* the long term memory is used to produce the note, and no information is passed on from one note to the next. In our case the reference pitch remains r^* throughout the piece, i.e. this simplifies to the constant reference pitch model given in Eq. (9). 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 Section VI.B) 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 ($F(67) = 30.51$, $p < 0.0001$). The increase in variance per note is 0.001, corresponding to an increase of 0.075 in variance over 75 notes, equivalent to a substantial increase in standard deviation of $\sqrt{0.075} = 0.27$. 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 pitch memory—predicts zero drift. The question is then whether a model with an intermediate memory value of $\mu \in (0, 1)$ will fit the data better.

D. An intermediate memory parameter μ

Having rejected the boundary models for $\mu = 0$ and $\mu = 1$ we are interested in finding whether any intermediate μ provides a more adequate model. A good model should predict

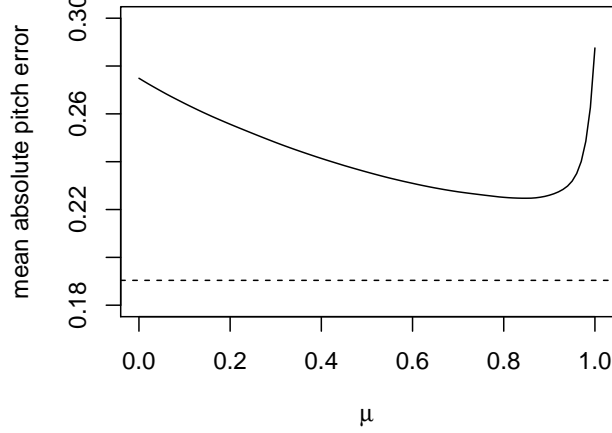


FIG. 9: Mean absolute error for models based on (11) for different values of the memory weight μ . An optimum is recognisable around $\mu = 0.85$. Dashed line: best linear prediction.

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 Figure 8, it seems plausible that, for some parameter μ the prediction error can become small. We measure the model’s mean absolute pitch error (*model MAPE*) with respect to this reference. Figure 9 shows the error on a grid of μ 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 towards 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 μ values that minimise the individual recordings and averaged them by singer to obtain singer-wise μ 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).

The model behaviour in both pitch prediction and spread of drift suggests that a memory model such as the one defined by Equations (10) and (11) is reasonable for values around

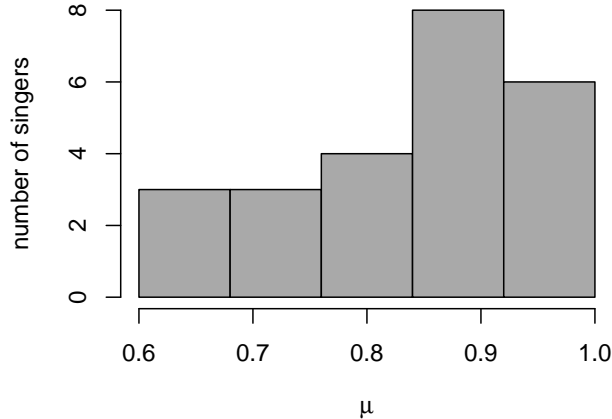


FIG. 10: Histogram of μ by singer.

$\mu = 0.85$.

VII. DISCUSSION

The intonation memory model presented above is particularly interesting because the parameter μ 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 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 our 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. For example, the melody contains notes from a single major scale, and only some of the intervals possible in that scale actually occur. More different melodies are needed to study intonation behaviour in more detail and with more claim to generality.

While we found that in our study equal temperament 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 observed error magnitudes in our experiments were

larger than typical differences between temperaments, so it is likely that such fine distinctions are more relevant to vertical harmony, where singers are able to tune to an external reference using the roughness of beating between partials of simultaneous notes.

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.

As we pointed out in the introduction, previous studies have dealt with intonation drift in polyphonic singing (Devaney and Ellis, 2008; Howard, 2007), and we deliberately studied the simpler case of unaccompanied solo singers. Much is to explore in between, especially interaction between singers; for example, investigating whether the process of inferring the reference intonation from another singer’s imperfect singing itself leads to biased intonation.

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, while evidently common, is often minor (less than 0.2 semitones over 50 notes), and not correlated to pitch accuracy, interval accuracy, or musical background. Surprisingly, most significant drifts are upward drifts. Using these findings on solo intonation drift we motivate a causal intonation memory model with a single parameter μ representing intonation memory strength. We show that values around $\mu = 0.85$ minimise the model mean absolute pitch error. Our discussion section highlights possibilities for future work, including further investigations of memory parameters on individuals, and a more diverse set of melodies.

IX. REFERENCES

References

- Barbershop Tuning Discussion (2012). “Barbershop tuning discussion”, URL <http://infohost.nmt.edu/~jstarret/bbshop2.html>, see supplementary material at [URL will be inserted by AIP] for a text transcript of the discussion.
- 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”, *Glott Int.* **5**, 341–345.
- Brown, D. (1991). *Human universals* (Temple University Press, Philadelphia).
- 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*, 1467–1468 (Firenze, Italy).
- 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*, 421–428.
- Crowther, D. S. (2003). *Key Choral Concepts: Teaching Techniques and Tools to Help Your Choir Sound Great!* (Horizon Publishers, Springville, Utah), 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., Giguere, 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. Interdiscipl. 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*, 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*, 219–224.
- Filzmoser, P., Garrett, R., and Reimann, C. (2005). “Multivariate outlier detection in exploration geochemistry”, *Computers & Geosciences* **13**, 579–587.
- Henning, G. B. (1955). “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.
- Kennedy, M. (1980). *The Concise Oxford Dictionary of Music* (Oxford University Press), p. 319.
- Lombard, E. (1911). “Le signe de l’élévation de la voix”, *Ann. Mal. Oreil. Larynx* **2**, 101–109.
- Markel, J. (1972). “The SIFT algorithm for fundamental frequency estimation”, *IEEE Trans. Audio and Electroacoust.* **20**, 367–377.
- Mithen, S. J. (2007). *The Singing Neanderthal: A Search for the Origins of Art, Religion, and Science* (Harvard University Press, Cambridge, Mass.), esp. Ch. 16, pp. 246–265.
- Molina, E. (2012). “Automatic scoring of singing voice based on melodic similarity measures”, Master’s thesis, Universitat Pompeu Fabra, pp. 9–14.
- Mürbe, D., Pabst, F., Hofmann, G., and Sundberg, J. (2002). “Significance of auditory and kinesthetic feedback to singers pitch control”, *J. of Voice* **16**, 44–51.
- Pfordresher, P. Q. and Brown, S. (2007). “Poor-pitch singing in the absence of “tone deafness””, *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.
- Pfordresher, P. Q. and Mantell, J. T. (2012). “Effects of altered auditory feedback across effector systems: Production of melodies by keyboard and singing”, Acta Psychol. **139**, 166–177.
- 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”, The Psychological Monographs **16**, 1–12.
- Seashore, C. E. (1967). *Psychology of music* (Dover Publications, New York), pp. 254–272.
- Swannell, J. (1992). *The Oxford Modern English Dictionary* (Oxford University Press, USA), p. 560.
- Team R Development Core (2008). *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria), URL <http://www.R-project.org>, visited August 2013.
- Vurma, A. and Ross, J. (2006). “Production and perception of musical intervals”, Music Percept. **23**, 331–344.

APPENDIX A: PFORDRESHER’S SINGING METRICS

We present the singing metrics of Pfordresher *et al.* (2010) and argue why we prefer alternative terminology and definitions for our study.

Note accuracy α_N is defined as the mean of signed deviations of sung pitches p_i from the target pitches p_i^0 :

$$\alpha_N = \frac{1}{M} \sum_{i=1}^M p_i - p_i^0.$$

Since the average is calculated from signed differences and is thus itself signed, deviations

in opposite directions could cancel each other. Thus the note accuracy measures a bias or systematic deviation to lower ($\alpha_N < 0$) or higher ($\alpha_N > 0$) pitches rather than accuracy per se. In unaccompanied singing, and hence in our study, there is no absolute reference pitch and hence this measure is not meaningful. Furthermore, as higher absolute values of α_N indicate *less* accuracy, we prefer the term *pitch bias* or *pitch offset*.

Note precision is defined as the mean standard deviation of pitches. Hence, if there are K pitch classes, with M_j instances of pitch class P_j having a mean μ_j , the variance s_j^2 for the pitch class is given by:

$$s_j^2 = \frac{1}{M_j - 1} \sum_{p_i \in P_j} (p_i - \mu_j)^2.$$

The note precision π_N is thus:

$$\pi_N = \frac{1}{K} \sum_{j=1}^K s_j.$$

As a mean standard deviation, note precision is unsigned and positive definite. The larger the value, the more dispersed are the sung pitches in each pitch class, the smaller the value, the more consistently the pitches are produced. Again an alternative term such as *pitch spread* would be more appropriate.

Interval accuracy is defined as the mean deviation of sung intervals $\Delta p_i = p_{i+1} - p_i$ from target intervals:

$$\alpha_I = \frac{1}{M - 1} \sum_{i=1}^{M-1} |\Delta p_i| - |\Delta p_i^0|.$$

Interval accuracy is itself signed. The sign indicates systematic deviations to smaller ($\alpha_I < 0$) or larger ($\alpha_I > 0$) intervals. As there is no distinction between ascending and descending intervals, two problems arise: the interval accuracy erroneously assesses an interval of the correct magnitude but wrong direction as being accurate; and a tendency to drift, for example to sing flat (downward pitch drift), is not captured, as this results in smaller ascending intervals but larger descending intervals, which cancel if the error magnitudes match.

Finally, *interval precision* is defined as the mean standard deviation of interval errors. Hence, if there are K interval classes I_j each having M_j instances with a mean of μ_j , the

variance s_j^2 is given by:

$$s_j^2 = \frac{1}{M_j - 1} \sum_{\Delta p_i \in I_j} (\Delta p_i - \mu_j)^2$$

and the interval precision π_I is thus:

$$\pi_I = \frac{1}{K} \sum_{j=1}^K s_j.$$

Once again, this is a measure of spread, with lower values indicating greater precision, so an alternative name such as *interval spread* would be preferable.

As a concrete example, consider a case where every interval of m semitones (in either direction) is sung m cents flat, the sum of all ascending intervals is n semitones, and the first and last note of the piece are the same nominal pitch. Then it can be shown that the interval accuracy $\alpha_I = 0$, the interval precision $\pi_I = 0$, but the piece has drifted by $2n$ cents downwards.

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

TABLE I: Self-reported musical experience.

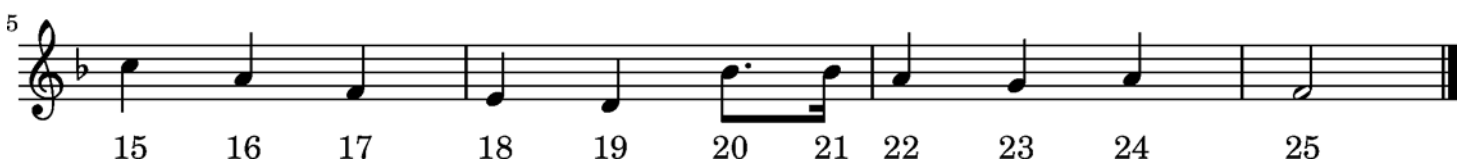
sg.abl	0.40	0.31	0.54	-0.45	-0.46	0.11	-0.02	0.06
	sg.exp	-0.07	0.42	-0.16	-0.27	0.20	0.05	0.11
		mus.bg	0.34	-0.16	-0.24	0.10	-0.02	0.05
		ch.exp	-0.37	-0.40	0.22	0.01	0.07	
			MAIE	0.93	-0.19	-0.01	-0.06	
				MAPE	-0.19	-0.01	-0.04	
					D_L	0.52	0.94	
						$ D $	0.54	
							D	

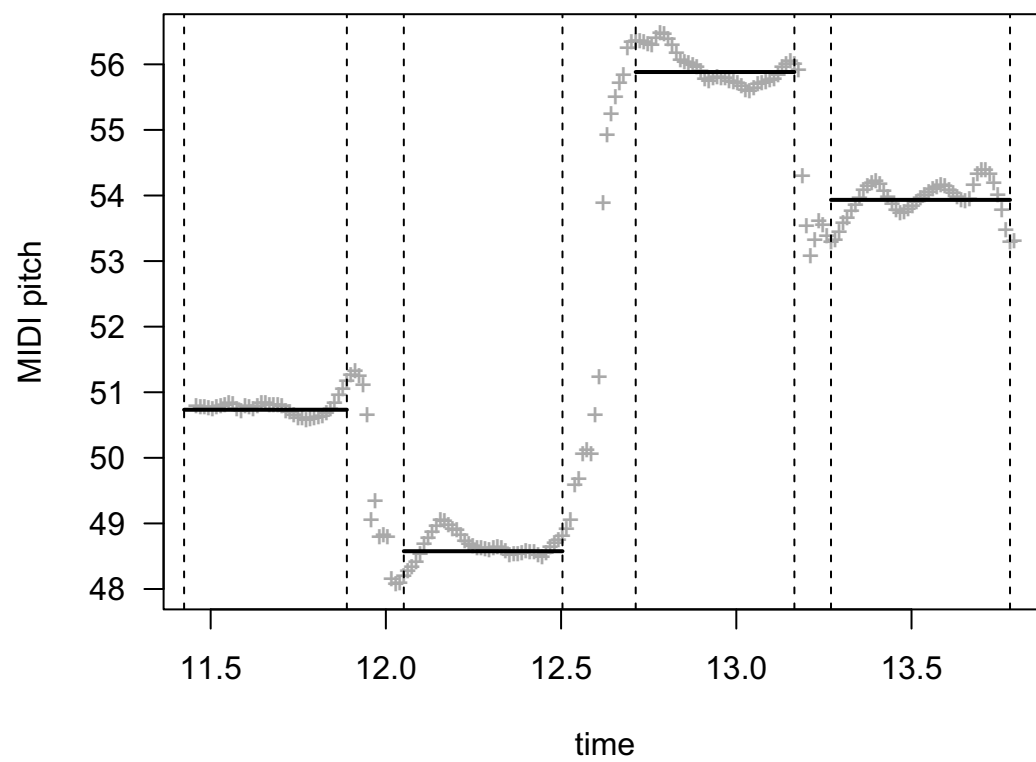
TABLE II: Spearman rank correlations of survey metadata (singing ability, singing experience, musical background, choir experience) and measures of accuracy and drift.

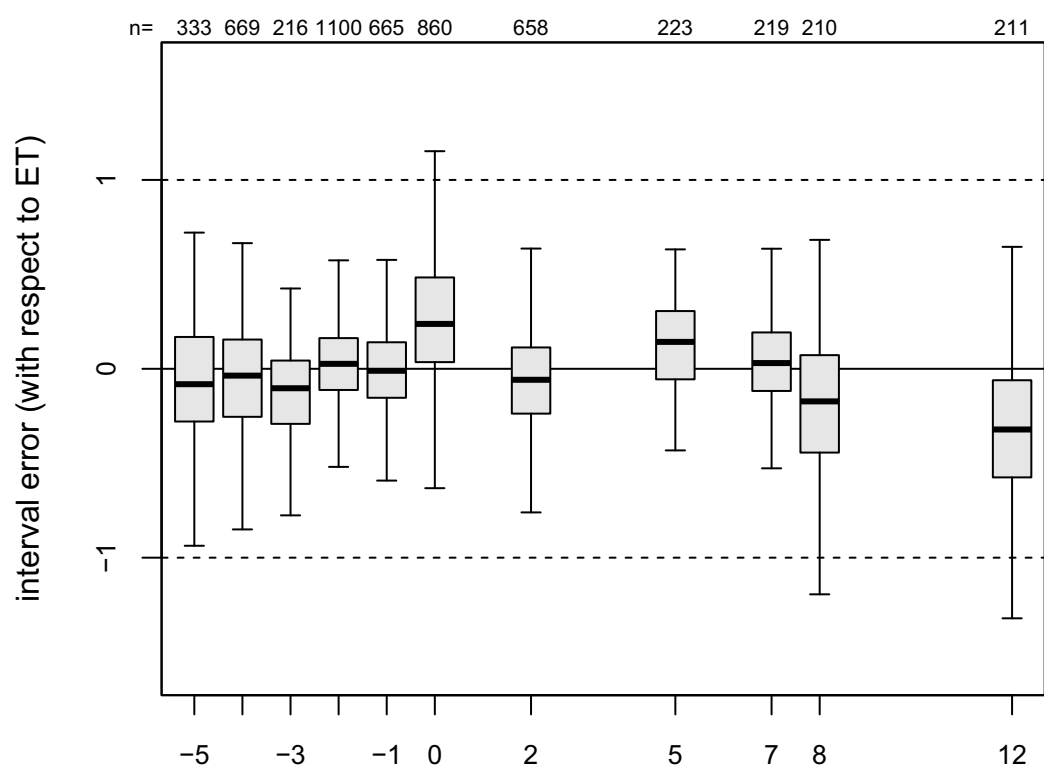
Significant correlations ($p < 0.01$) are shown in bold.

LIST OF FIGURES

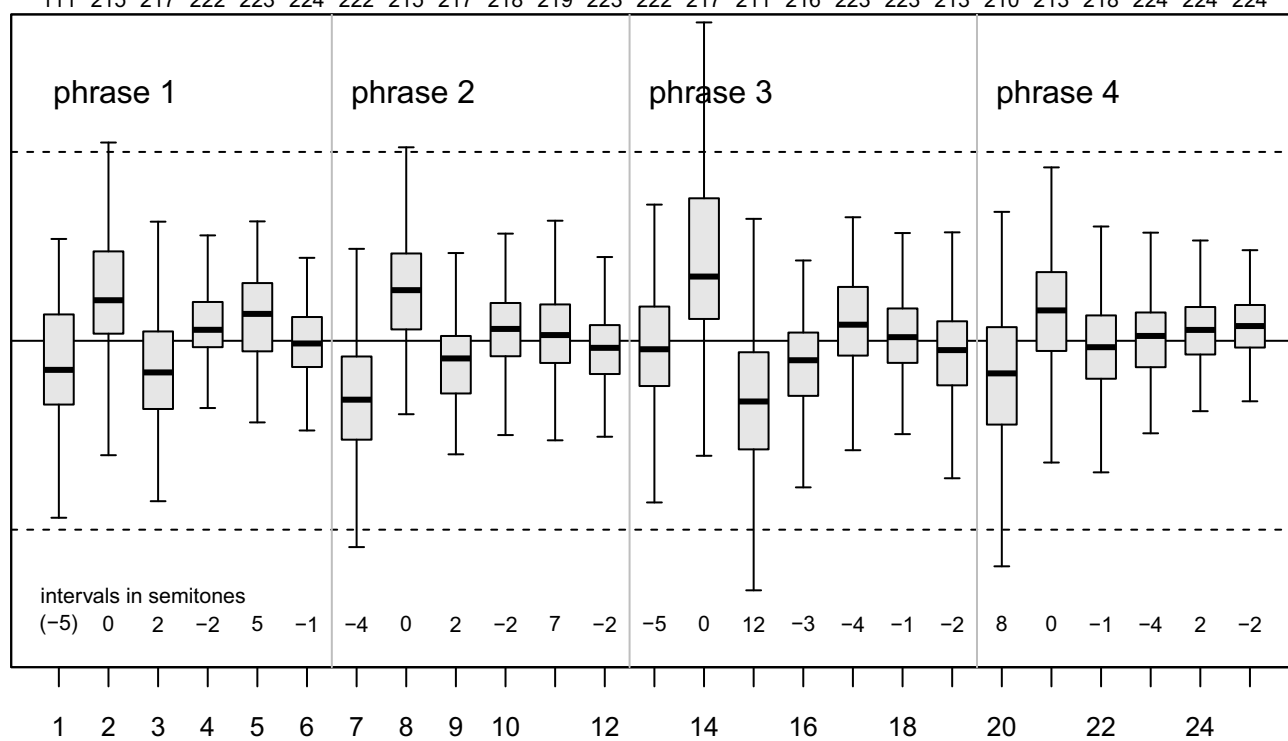
FIG. 1	“Happy Birthday” in F-Major	7
FIG. 2	Example pitch track (grey crosses) and note-wise pitch estimates (horizontal bars), calculated as medians between annotated note boundaries (vertical dashed lines).	10
FIG. 3	Interval errors in semitones relative to the score, using equal temperament. The boxes indicate the 1st, 2nd (median) and 3rd quartiles, the whiskers extend to ‘the most extreme data point which is no more than 1.5 times the interquartile range’ (R software Team R Development Core (2008)). Outliers are omitted for clarity of display.	11
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	12
FIG. 5	Pitch errors with respect to linear prediction (run-wise).	13
FIG. 6	Distributions of singing accuracy metrics over all conditions and participants.	16
FIG. 7	Significance of drift, showing p values (logarithmic scale) against D_L for each recording. p values below 0.01 are considered significant.	18
FIG. 8	Example of observed tonality estimates t_i (marked as +) and the estimated reference pitch r_i (filled bullets) with parameter $\mu = 0.85$	23
FIG. 9	Mean absolute error for models based on (11) for different values of the memory weight μ . An optimum is recognisable around $\mu = 0.85$. Dashed line: best linear prediction.	25
FIG. 10	Histogram of μ by singer.	26

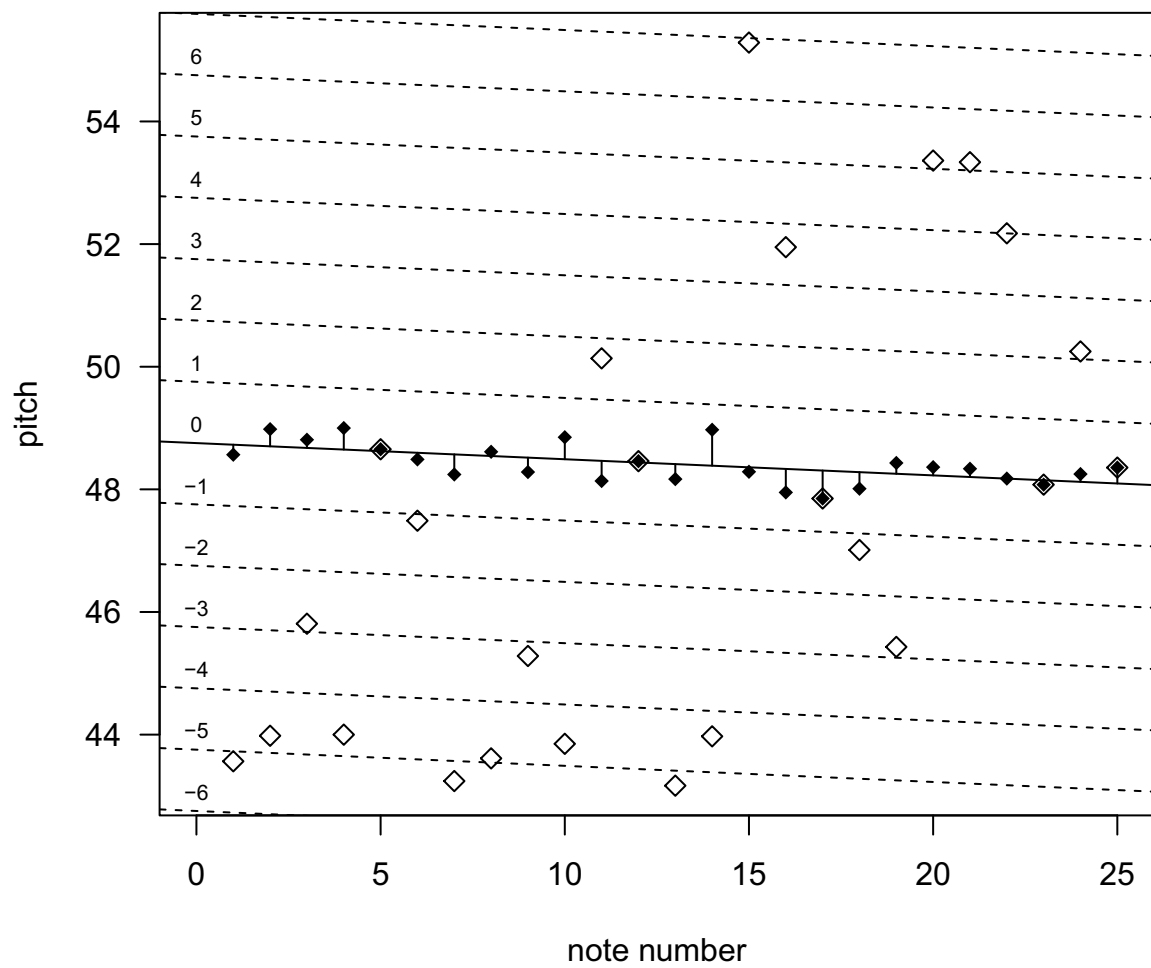






111 215 217 222 223 224 222 215 217 218 219 223 222 217 211 216 223 223 213 210 213 218 224 224 224





pitch error (with respect to ET)

