The Journal of the Acoustical Society of America

Singing Together: Pitch Accuracy and Interaction in Unaccompanied Duet Singing

--Manuscript Draft--

Manuscript Number:	JASA-03137R2
Full Title:	Singing Together: Pitch Accuracy and Interaction in Unaccompanied Duet Singing
Short Title:	Pitch Accuracy and Interaction in Singing
Article Type:	Regular Article
Corresponding Author:	Jiajie Dai, Ph.D. Queen Mary University of London London, UNITED KINGDOM
First Author:	Jiajie Dai, Ph.D.
Order of Authors:	Jiajie Dai, Ph.D.
	Simon Dixon, Professor
Section/Category:	Musical Acoustics
Keywords:	unaccompanied singing; pitch accuracy; pitch drift; interaction in duet.
Abstract:	We investigate singing interaction by analysis of the factors influencing pitch accuracy of unaccompanied duet singers. Eight pairs of singers sang two excerpts either in unison or two-part harmony. The experimental condition varied which singers could hear their partners. After semi-automatic pitch-tracking and manual checking, we calculated the pitch error and interval error, and tested the factors of influence using a one-way ANOVA. The results indicate that: 1) singing with the same vocal part is more accurate than singing with a different vocal part; 2) singing solo has less pitch error than singing with a partner; 3) pitch errors are correlated, as singers adjust their pitch to mitigate their partner's error and preserve harmonic intervals at the expense of melodic intervals and absolute pitch; 4) other factors influence the pitch accuracy, including: score pitch, score harmonic interval, score melodic interval, musical background, vocal part and individual differences.

Singing Together: Pitch Accuracy and Interaction in Unaccompanied Unison and Duet Singing

Jiajie Dai[1, a)] and Simon Dixon[1, b]

Centre for Digital Music, Queen Mary University of London

(Dated: 16 December 2018)

We investigate singing interaction by analysis of the factors influencing pitch accuracy of unaccompanied pairs of singers. Eight pairs of singers sang two excerpts either in unison or two-part harmony. The experimental condition varied which singers could hear their partners. After semi-automatic pitch-tracking and manual checking, we calculated the pitch error and interval error, and tested the factors of influence using a one-way ANOVA and a linear mixed-effects model. The results indicate that: 1) singing with the same vocal part is more accurate than singing with a different vocal part; 2) singing solo has less pitch error than singing with a partner; 3) pitch errors are correlated, as singers adjust their pitch to mitigate their partner's error and preserve harmonic intervals at the expense of melodic intervals and absolute pitch; 4) other factors influence the pitch accuracy, including: score pitch, score harmonic interval, score melodic interval, musical background, vocal part and individual differences.

PACS numbers: 43.75.Rs, 43.75.Bc, 43.75.Xz, 43.75.St

---

a) j.dai@qmul.ac.uk

b) s.e.dixon@qmul.ac.uk

I. INTRODUCTION

Singing is common to all human societies (Brown, 1991) and repertoire performed by multiple singers is probably the most widespread type of singing (Sundberg, 1987), yet the factors that affect the accuracy of group singing are still poorly understood. The main motivation for this study is to improve the scientific understanding of unaccompanied duet singing, and in particular the interaction between singers. We seek to explain pitch accuracy and the mechanisms which may influence tuning in complex situations. The basic concepts of pitch accuracy and interaction are introduced in this section and relevant research in the next section.

Intonation in music is defined as a musician's realisation of pitch accuracy (Simpson *et al.*, 1989). It is one of the central parameters of singing accuracy and it is an extremely significant aspect of music because of its relevance to both melody and harmony. The accuracy of intonation is determined by culturally specific tuning systems such as the equal tempered tuning system in Western music (Warren and Curtis, 2015). Intonation is the main reported priority in choral rehearsals (Ganschow, 2014) and the focus of guides on vocal practice (Crowther, 2003).

To produce an accurate pitch, most people rely on a recent reference (Takeuchi and Hulse, 1993). Therefore, the accompaniment of instruments and other singers, where present, plays an important role in tuning. Although instrumental accompaniment has been shown to enhance individual learning of a piece (Brandler and Peynircioglu, 2015), it can also reduce pitch accuracy during singing, even when the accompaniment consists of nothing but the target pitches (Dai and Dixon, 2016; Pfordresher and Brown, 2007).

In the case of fixed pitch instruments, such as keyboard instruments, singers adjust to the tonal reference provided by the instrument. But in unaccompanied singing, the singers negotiate a common reference, and this reference can change over time. Several studies have investigated the intonation of unaccompanied ensembles and how their tonal reference evolves over the duration of a piece, a phenomenon called *pitch drift* (see Section II). Alldahl (2008) cites relative pitches, singers' memories and their muscle control as critical factors influencing intonation, but little is known about the effect of interaction between singers.

Interaction is very important for ensemble singing, which is a cooperative activity involving communication within the ensemble and with the audience (Potter, 2000, p. 158). Attaining excellence in ensemble playing depends on finding a balance between individual performance and interaction (Lim, 2014). This research investigates how singers influence each other in terms of intonation and pitch variation. We focus on duet singing as the simplest example of singing involving interaction, allowing us to design a controlled experiment involving the influence of one singer upon another.

The remainder of the paper is structured as follows. Section II discusses existing work related to singing intonation and interaction. Section III contains our research questions, hypotheses, experimental design and methodology. In Section IV, we describe our data analysis, including annotation and calculation of intonation metrics. Section V presents our results and how they relate to the experimental hypotheses. The combined effect of multiple factors is evaluated in a linear mixed effects model in Section VI. This is followed by a discussion of the results (Section VII), our conclusions (Section VIII), and finally the details of where the annotated data and software can be freely obtained (Section IX).

II. PREVIOUS WORK

Research quantifying the intonation of vocal sounds can be traced back over 100 years to the early work of Seashore (1914), and continues until the present time. Pitch production relies on the ability to control the tension in the vocal cords, which results in modulations of the vocal fundamental frequency. Much vocal research has focussed on speech, but musical pitch requires a much greater degree of accuracy, both in production and perception, than speech (Zatorre and Baum, 2012). Abilities related to the control of pitch are the primary indicator for distinguishing untrained but talented individuals from those with less innate singing skills (Watts *et al.*, 2003).

In order to study intonation in audio recordings, a reliable pitch estimation algorithm is required. Note that since the voiced part of vocal sounds is harmonic, pitch and fundamental frequency (f_0) are generally treated as exchangeable (although they are expressed on different scales, Equation 1). Many pitch detection methods have been proposed, particularly for speech recognition and coding (e.g. Gerhard, 2003; Hess, 1983; Rabiner *et al.*, 1976). If only a single pitch is present in the signal, periodicity-based methods such as autocorrelation, as in the widely used Praat system (Boersma, 2002), and difference functions, as in YIN (de Cheveigné and Kawahara, 2002), are popular approaches for determining the pitch of speech or musical sounds. In this work we use PYIN (Mauch and Dixon, 2014), a probabilistic extension of YIN which provides robustness against errors due to suboptimal threshold settings.

Most studies on intonation focus on accuracy, although topics such as vibrato have also been investigated (Bretos and Sundberg, 2003; Ferrante, 2011). Note that we use "accuracy" to refer to both the bias and spread of pitch errors (unlike Pfordresher and Brown (2007), who use it specifically for the bias alone). On the one hand, pitch error is the main metric of accuracy for many researchers, where each observed pitch is compared to a predetermined target value. Several studies have investigated pitch drift in unaccompanied singing (e.g. Devaney and Ellis, 2008; Howard, 2003; Kalin, 2005; Mauch *et al.*, 2014; Terasawa, 2004). Howard (2007) tested the hypothesis that the use of *just intonation*, where the fundamental frequencies of pairs of simultaneous or consecutive notes are related by ratios of small whole numbers (Lindley, 2001), causes pitch drift. The hypothesis in such work is that the pitch adjustments required to intone pure intervals accumulate over time resulting in a shifting tonal reference (Mullen, 2000). Howard's study confirmed that singers make use of non-equal-tempered intonation to govern their tuning, and showed that it is possible to predict the direction of pitch drift in controlled harmonic progressions.

On the other hand, interval error, the extent to which pitch differences between subsequent tones deviate from their target values, has also been investigated. Tritones (Dai *et al.*, 2015) and perfect fifths (Vurma and Ross, 2006) were reported to have greater interval error than other intervals. Other authors observed a phenomenon called *compression*, whereby sung intervals are smaller than their targets, an effect which is particularly strong amongst unskilled singers (Pfordresher and Brown, 2007).

Individual factors such as age and sex influence pitch accuracy (Welch *et al.*, 1997). Musical training and experience also have some influence on singing ability; Mauch *et al.* (2014) found that self-rated singing ability and choir experience, but not general musical background, correlated significantly with intonation accuracy. Singers who exhibit much greater than average pitch errors are classified as *poor singers*, a phenomenon that has been the focus of several studies (Berkowska and Dalla Bella, 2009; Dalla Bella *et al.*, 2007; Pfordresher and Brown, 2007; Pfordresher *et al.*, 2010). For poor pitch singing, evidence points to a deficiency in pitch imitation accuracy as the main cause (Pfordresher and Mantell, 2014), although there are several types of singing deficiency and they vary by age and training (e.g. Demorest *et al.*, 2015).

Mürbe *et al.* (2002) showed how singers' intonation accuracy is reduced by diminished auditory feedback; in their experiment, auditory feedback was masked by noise. When singers cannot hear themselves, they have to rely on kinesthetic feedback circuits, which are less effective than auditory feedback for informing intonation. Likewise even in musical situations where the accompanying sound provides the tonal reference, singers make greater pitch errors when singing with accompaniment (Pfordresher and Brown, 2007), and particularly when the accompanying pitch content varies over the duration of a note (Dai and Dixon, 2016). Thus vocal accompaniment is more difficult to sing with than instrumental accompaniment, because singers are relying on unstable reference pitches from other vocal parts (Liimola, 2000, p. 151). Although singing in unison with a partner may not increase pitch accuracy, it may give singers more confidence than singing solo (Heath and Gonzalez, 1995).

Previous studies have investigated differences between solo and unison singing, although not all studies obtained significant results. For example, Green (1994) claimed that children singing unison, as opposed to in individually, had significantly better vocal accuracy, while Cooper (1995) was unable to show a significant difference. There are more observations also show children sing more accurately individually than in a group ((e.g. Clayton, 1986; Goetze, 1985, 1989)). Besides the singing conditions, age, gender, training and number of attempts were reported as significant factors for children's singing accuracy ((e.g. Nichols, 2016; Nichols and Wang, 2016)).

Except for the 0.01% of the population who have absolute pitch, the ability to identify or reproduce any given pitch on demand (Bohrer, 2002; Takeuchi and Hulse, 1993), most people rely on a reference pitch for tuning. An initial reference will be forgotten over time (Long, 1977; Mauch *et al.*, 2014), so singers must constantly update their frame of reference as they sing, based on what they have recently heard, both their own voice and any accompaniment.

Brandler and Peynircioglu (2015) observed that participants learned new pieces of music more successfully when in an individual learning environment than in a collaborative one. Abundant evidence shows that singers are influenced by other choral members in terms of pitch accuracy (e.g. Howard, 2003; Terasawa, 2004) and various approaches have been proposed to keep singers in tune by their relative pitches, tone memories and muscle memories (e.g. Alldahl, 2008; Bohrer, 2002). Although various studies on singing have investigated the pitch accuracy of solo singers and singing ensembles, we are not aware of any work that focusses directly on the interaction between singers and its effect on intonation, the topic of this study.

III. METHODOLOGY

In this section, we describe our hypotheses, the experimental design, musical material, participants and experimental procedure. For our experiment, two *singing conditions* are defined: the *unison condition*, where two singers sing the same vocal part, and the *duet condition*, where they sing different vocal parts. There are also four *listening conditions.* In the *solo* condition, the two singers cannot hear each other. The two *simplex* conditions are where only one singer can hear the other singer (in either direction). The singer who cannot hear her partner is called the *independent singer* while the singer who hears her partner is the *dependent singer.* The *duplex* condition is where both singers can hear each other. Note that according to these definitions, both singers are independent in the solo condition, and both are dependent in the duplex condition. Singers can hear their own voice in all conditions.

A. Hypotheses

Based on previous research and musical experience, we formulated five hypotheses regarding effects we expected to observe when singers interact. The experimental method was designed to test these hypotheses and quantify the extent of the effects observed.

Hypothesis 1: *The unison singing condition has less pitch error, melodic and harmonic interval error than the duet condition.* Participants sing the same pitch in the unison singing condition while they sing harmony in the duet condition. An observation from choral singing is that most singers, particularly those with less musical training, find it easier to sing their vocal part when others around them are singing the same part. Singing in harmony with different parts requires greater concentration, to avoid being distracted from one's own part.

Hypothesis 2: *Independent singers have less pitch error than dependent singers.* Auditory feedback is essential for accurate intonation. As either noise (Mürbe *et al.*, 2002) or simultaneously playing the target melody (Dai and Dixon, 2016; Pfordresher and Brown, 2007) reduces singers' accuracy, we expect to observe this effect in both singing conditions. Although comparisons of pitch accuracy in unison versus solo singing did not always agree with each other, the majority of existing evidence suggests that individual singing is more accurate than unison singing (e.g. Clayton, 1986; Goetze, 1985, 1989).

Hypothesis 3: *The duplex condition has less harmonic interval error than the solo condition.* When singers do not hear each other, their errors are independent as it is impossible for them to adjust their intervals according to their partner's intonation. When they can hear their partner, they adjust their pitch in order to reduce the harmonic interval error. Since most of the singers have choral experience, this hypothesis is based on the assumption that such singers are somewhat able to attune to other singers and sing harmoniously as a group, which is an important skill that is practised in their rehearsals (Bohrer, 2002).

Hypothesis 4: *There is a positive correlation between the pitch error of the dependent singer and the independent singer in the simplex conditions.* The simplex condition allows for a one-way influence of the intonation of the independent singer upon the dependent singer. We predict that this influence will be seen not only in the magnitude of pitch errors (it is harder to sing well when distracted by an out of tune partner), but also in the direction of these errors (the dependent singer will adjust their pitch to reduce errors in vertical harmonies at the expense of absolute pitch error and melodic interval error). Thus a significant correlation between the pitch errors of dependent and independent singers provides evidence of interaction. Although features of the score could explain correlation in the unison condition (e.g. where both singers compress leaps), we predict this effect to hold also for the duet condition, where the score would not have a uniform effect on both singers.

Hypothesis 5: *The within-note pitch variation of dependent singers is higher than that of independent singers.* Our final hypothesis relates to the variation of pitch within each tone, which provides another view of interaction between singers. In the independent condition, any adjustment of pitch within a note arises from the singer's own feedback loop and involuntary noise in the vocal production system. In the dependent condition, there is also scope for intentional adjustment to improve harmonic intervals, as well as unintentional changes due to the distraction of hearing another singer.

B. Design

To test these hypotheses, we designed and implemented a controlled experiment involving two musical excerpts, two singing conditions (unison and duet) and three types of listening conditions (solo, simplex, duplex), as listed in Table I. Each trial involves two singers, denoted A and B. In the unison condition both singers sing the same vocal part (either the soprano or alto part). In the duet condition, singer A sings the soprano part and singer B the alto. For the listening conditions, the solo condition acts as a control, where the two singers sing separately without hearing each other. In the two simplex conditions, only one singer can hear their partner, with the direction of auditory feedback being reversed

Singing Condition	Listening Condition	A sings	B sings	A hears B	B hears A
Unison	Solo	Soprano	Soprano	No	No
Unison	Simplex	Soprano	Soprano	Yes	No
Unison	Simplex	Soprano	Soprano	No	Yes
Unison	Duplex	Soprano	Soprano	Yes	Yes
Unison	Solo	Alto	Alto	No	No
Unison	Simplex	Alto	Alto	Yes	No
Unison	Simplex	Alto	Alto	No	Yes
Unison	Duplex	Alto	Alto	Yes	Yes
Duet	Solo	Soprano	Alto	No	No
Duet	Simplex	Soprano	Alto	Yes	No
Duet	Simplex	Soprano	Alto	No	Yes
Duet	Duplex	Soprano	Alto	Yes	Yes

TABLE I. Experimental design for two singers A and B: singing and listening conditions.

between the two conditions. Finally in the duplex condition, both singers hear the voice of their partner. Except for the voice of their partner in certain listening conditions, there is no accompaniment during the experiment.

C. Musical Materials

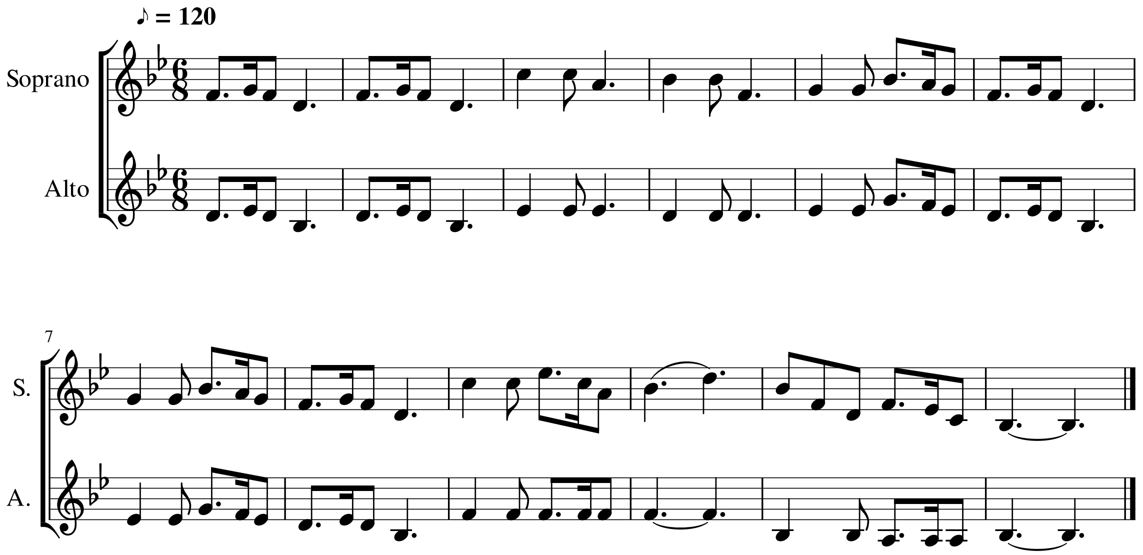
We chose the soprano and alto parts of two common choral pieces "Silent Night" (Gruber, c.1816) and "O Sacred Head, Now Wounded" (melody by Hassler, c.1601, harmonised by J.S. Bach, c.1729) as our experimental materials. These two pieces are examples of the traditional Western church choir repertoire with the former song being particularly well-known. The pitch range is from A3 to E♭5 (soprano: B♭3 to E♭5; alto: A3 to G4) with various melodic and harmonic intervals up to a minor 7th. The second piece was shortened to its first 12 bars as shown in Figure 1 to match the lengths of the two pieces.

D. Participants

Although factors of age and gender affect pitch accuracy (Welch *et al.*, 1997), they are not a target of this research. As our musical material consisted of soprano and alto parts, we recruited female singers only. Because this experiment required singers to maintain their own part while the other singer sang a different part, we recruited participants who have choral experience. All participants are amateur singers who have some musical training, and are members of our university's music society, *a capella* society or our research group. Pairs were allocated according to voice (one soprano, one alto) and availability. Although some sing together in the same choir, no pair had sung together in a duet or small group before the experiment. Each participant was involved in only one pair.

16 female UK residents took part in this experiment, with an age range from 19 to 30 years old (mean: 23.1; median: 23.5; SD: 3.3). Eight of the participants identified

Silent Night

John F. Young 1863 Franz X. Gruber circa 1816-1818

Piece 1: Silent Night

O Sacred Head, Now Wounded

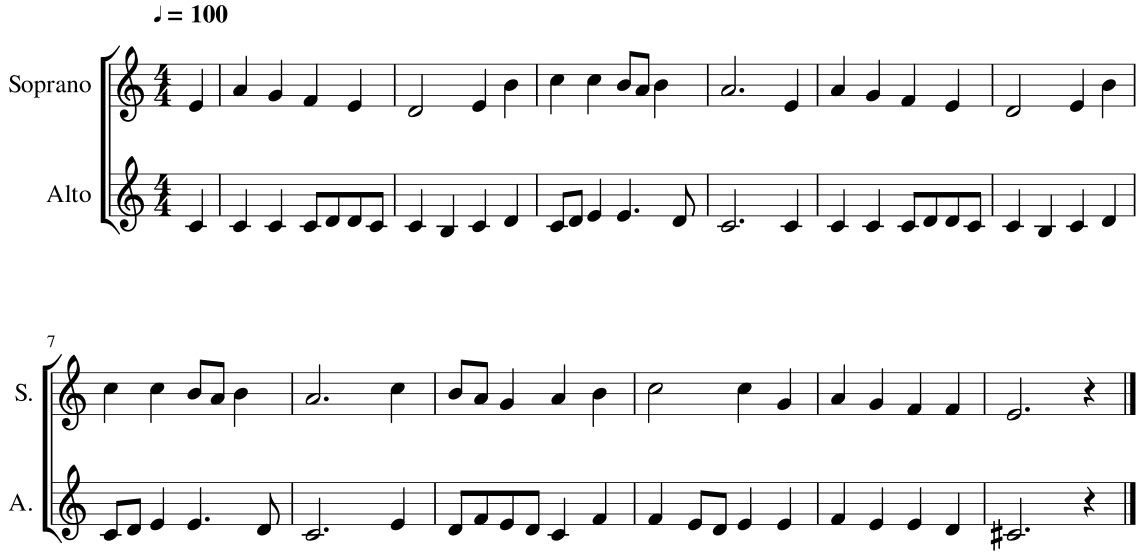
James W. Alexander, 1830 Adapted by J. S. Bach 1729

Piece 2: O Sacred Head, Now Wounded (first 12 bars)

FIG. 1. Musical material selected for the experiments.

themselves as sopranos, the other eight as altos. The sopranos (age range: 19–27; mean: 23.0; median: 24.0; SD: 3.0) and altos (age range: 19–30; mean: 23.3; median: 22.5; SD: 3.4) had similar age distributions. All the participants were able to sing the pitch range from A3 to E♭5 naturally, and could sing both pieces independently. In order to identify and exclude any poor singers (Pfordresher and Brown, 2007), we calculated the mean absolute melodic interval error (Equation 6) of each singer and planned to exclude any with an error greater than 0.5 semitones; no singer needed to be excluded.

For testing the effect of training, all the participants completed a self-assessment questionnaire based on the Goldsmiths Musical Sophistication Index (Müllensiefen *et al.*, 2014) which can be grouped into 4 main factors for analysis: active engagement, perceptual abilities, musical training and singing ability (9, 9, 7 and 7 questions respectively). The proportion of singers having more than three years of choir experience is 62.5%; all have at least one year of instrumental training; and 50.0% of the participants have at least six years of formal training on musical instrument or voice.

E. Procedure

The study was conducted with the approval of the Queen Mary Ethics of Research Committee (approval number: QMREC1456). The participants were grouped into eight pairs of singers, each consisting of one soprano (singer A) and one alto (singer B) by self-identification. Each pair participated in both the unison and duet singing conditions. Each singer sang the two pieces in each of the four listening conditions as a set of data, resulting in eight pairs of duet datasets, eight pairs of unison soprano and eight pairs of unison alto datasets collected in this experiment, each consisting of eight recordings. All 384 recordings were grouped and labelled with the pair number, music piece, experimental conditions and the singer's questionnaire results for analysis.

Before the recording, the singers were given about half an hour to warm up and become familiar with the pieces. Participants practised their vocal parts with piano and their partners. The recording did not start until the participants could sing their vocal parts individually while their partner was singing the other part. At the beginning of each trial, participants heard instructions identifying the piece and condition and were given their own starting pitch repeated four times on a digital piano. During each trial, singers could hear a metronome and read the music score, but no further reference pitch was provided, nor did the participants talk to each other until the trial was completed. The trials were recorded in the same order with the same equipment (described below). To avoid any effect of vowel sound, and to assist annotation of note onset times, the participants were asked to sing the syllable /ta:/ rather than the lyrics. The participants could not see their partner during the trials. The total time of the experiment, including rehearsal, four listening conditions and questionnaire, was about one and a half hours.

The experiment was performed in two acoustically isolated rooms at the authors' university with facilities for multi-track recording (Morrell *et al.*, 2011). The equipment included an SSL MADI-AX analogue to digital converter, two Shure SM58 microphones and sound isolating headphones (Beyer Dynamic DT100). All the tracks were controlled and recorded with the software Logic Pro 10. The metronome and the reference pitches were also given by Logic Pro. The two microphone signals and (for reference) the two headphone signals were recorded on four separate tracks with a sampling rate of 44100 Hz and stored in .wav format. The total latency of the system is 4.9 ms from microphone to headphone, where 3.3 ms is due to the processing time of Logic Pro and 1.6 ms (71/44100) due to the converter.

IV. DATA ANALYSIS

This section describes the annotation procedure and the measurement of the four metrics of accuracy (pitch error, melodic interval error, harmonic interval error and pitch variation; defined below). These metrics are the dependent variables for hypothesis testing, while test and listening conditions are the main independent variables.

A. Annotation

We used the software *Tony* (Mauch *et al.*, 2015) to annotate the recordings with fundamental frequencies as extracted by the PYIN algorithm (Mauch and Dixon, 2014). The *Tony* software segments the recording into notes and silences, and outputs the median fundamental frequency f_0 for each note. The conversion of fundamental frequency to musical pitch p is calculated as follows:

$$\mathsf{p} = 69 + 12\log_2 \frac{\mathsf{f}_0}{440}. \tag{1}$$

This scale is chosen such that its units are semitones, with integer values of p coinciding with MIDI pitch numbers, and reference pitch A4 ($\mathsf{p} = 69$) tuned to 440 Hz. After automatic annotation, every single note was checked manually by the first author to make sure the tracking was consistent with the data and corrected if it was not. The annotation of all 384 files took over 31 hours, and resulted in a database of 18176 annotated notes (2 singers × 2 pieces × 4 trials × (1 duet + 2 unison) × 8 groups = 384 files).

The information in our database includes: group number, singer number, singing condition, listening condition, piece number, note in trial, score onset position, score duration, score pitch, score interval, observed onset time, observed duration, observed pitch, pitch error, melodic interval error, harmonic interval error, anonymised participant details, and questionnaire scores. We also store the pitch trajectory for each note. The data will be published for subsequent research (Section IX).

B. Metrics of Accuracy

Our metrics of intonation accuracy are pitch error, interval error, and pitch variation, defined below. The definitions of pitch error and interval error are based on Dai and Dixon (2017); Mauch *et al.* (2014), while pitch variability is inspired by Pfordresher *et al.* (2010).

1. Pitch Error

Pitch error e_i^p for note i is the difference between the observed pitch and score pitch:

$$e_i^p = \bar{p}_i - p_i^s, \tag{2}$$

where $\bar{p}_i$ is the median of the observed pitch trajectory of note i (calculated over the duration of an individual note), and p_i^s is the score pitch of note i as defined by the MIDI standard, where pitches are indexed by the note number from the beginning of the piece.

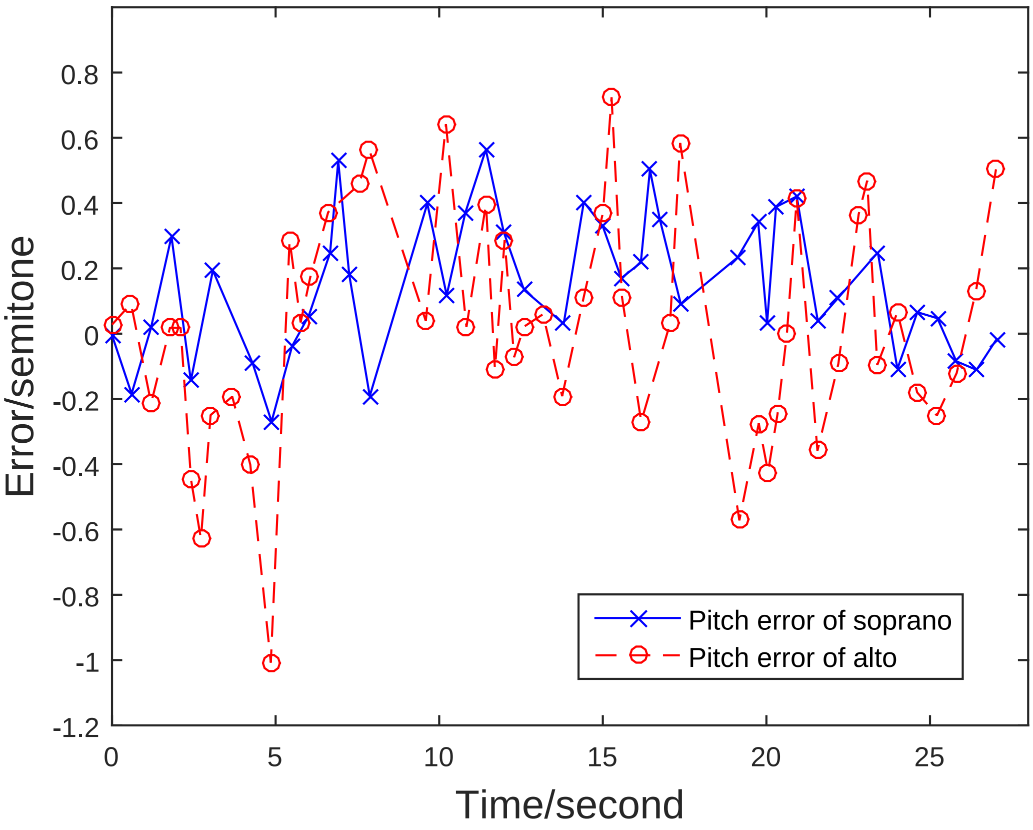


FIG. 2. Example of pitch error for piece 2, duet singing condition, duplex listening condition, for one pair of singers.

For example, when someone sings a score pitch of C5 at 510.34 Hz, this corresponds to $p = 71.57$ semitones (Equation 1), whereas the nominal pitch of C5 is 72. So the pitch error is $e^p = 71.57 - 72 = -0.43$ semitones. Pitch error measures the cumulative intonation error relative to the given starting tone. Figure 2 shows an example of pitch error for two singers in the duplex duet condition.

2. Interval Error

A musical interval is the difference between two pitches (Prout, 2011), which is proportional to the logarithm of the ratio of the corresponding fundamental frequencies. We distinguish two types of interval: a *melodic interval* is the pitch difference between two successive notes from a single singer, and a *harmonic interval* is the pitch difference between two simultaneous notes from different singers.

We define the melodic interval error e_i^m between the ith sung interval and the corresponding score interval as:

$$e_i^m = (\bar{p}_{i+1} - \bar{p}_i) - (p_{i+1}^s - p_i^s), \tag{3}$$

For example, if F4 is sung at $\bar{p}_i = 65.74$ and the subsequent note C5 at $\bar{p}_{i+1} = 71.57$, there should be a difference of $72-65 = 7$ semitones, but the observed difference is 5.83 semitones. So the melodic interval error for this case is -1.17 semitones.

The harmonic interval error is defined similarly: we subtract the score interval from the observed harmonic interval, as in equation 3. The notation is more complex in this case as: (1) a subscript is added to identify the singers; and (2) simultaneous notes might not always share the same sequence index, due to rests or multiple notes in one part while there is a single note in the other. The harmonic interval error e_k^h between singers A and B is:

$$e_k^h = (\bar{p}_{A,i} - \bar{p}_{B,j}) - (p_{A,i}^s - p_{B,j}^s), \tag{4}$$

where $p_{x,y}$ is the yth pitch of singer x, with $\bar{p}$ and p^s used as above, and notes (A, i) and (B, j) are assumed to be simultaneous (or at least overlapping in time).

Pitch error measures the absolute tuning, while melodic interval error captures local tuning within a vocal part. Harmonic interval error captures the local tuning between vocal parts, thereby facilitating analysis of the interaction between two singers.

3. *Pitch Accuracy over Multiple Notes*

To evaluate the pitch accuracy over a group of notes, we use the mean absolute value of each type of error as a summary measurement. For a group of M notes with pitch errors $\{e_1^p, \ldots, e_M^p\}$, the *mean absolute pitch error* (MAPE) is defined as:

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

The *mean absolute melodic interval error* (MAMIE) over M intervals is given by:

$$\text{MAMIE} = \frac{1}{M}\sum_{i=1}^{M} |e_i^m|, \tag{6}$$

and the *mean absolute harmonic interval error* (MAHIE) is defined similarly as:

$$\text{MAHIE} = \frac{1}{M}\sum_{i=1}^{M} |e_i^h|. \tag{7}$$

4. *Pitch Variation*

The pitch variation of a note is defined as the mean square pitch difference of the note trajectory from its median value. It indicates the extent of pitch variation over the duration of the note. The larger the pitch variation, the less stable the pitch. For a single note with N sampling points, where $p(i)$ represents the pitch at sampling point i and $\bar{p}$ is the median of $p(i)$ over the N points, the pitch variation V is calculated as follows:

$$V = \frac{1}{N}\sum_{i=1}^{N} |p(i) - \bar{p}|^2, \tag{8}$$

where the default sampling period for *Tony* is 5.8 ms. The *mean pitch variation* (MPV) is the mean value of pitch variation over multiple notes.

V. RESULTS

We calculated MAPE (Equation 5), MAMIE (Equation 6), MAHIE (Equation 7) and pitch variation (Equation 8) for each condition. In addition to the experimental conditions, we tested other possible factors for their effect on singing intonation. Over all conditions, the singers had an MAPE of 36 cents (SD=39), MAMIE of 24 cents (SD=28) and MAHIE of 41 cents (SD=47). We grouped the MAPE according to different factors, and fitted the grouped data separately into a one-way analysis of variance (ANOVA) model for testing the influence of each individual factor. The ANOVAs showed that the following factors influence the MAPE and MAMIE : singing condition, listening condition, score pitch, score melodic interval, score harmonic interval, note duration, piece, vocal part, singer, age and musical background (Table II). As harmonic intervals involve notes from both singers, MAHIE cannot test factors such as score pitch and vocal part. The ANOVA showed that singing condition, listening condition, note number in trial, music piece and score harmonic interval have a significant effect on MAHIE.

In this section, we focus on single factors of influence to test our hypotheses concerning intonation accuracy and pitch variation across the various experimental conditions.

Factor	**MAPE**	**MAMIE**	**MAHIE**
Singing condition	$F(1, 18174) = 70.8$ ***	$F(1, 18174) = 17.0$ ***	$F(1, 9086) = 316.7$ ***
Listening condition	$F(3, 18172) = 52.2$ ***	$F(3, 18172) = 41.0$ ***	$F(3, 9084) = 16.1$ ***
Note number in trial	$F(54, 18121) = 6.4$ ***	$F(54, 18121) = 15.2$ ***	$F(54, 9033) = 1.8$ ***
Score pitch	$F(15, 17552) = 22.3$ ***	$F(15, 17552) = 12.7$ ***	
Score melodic interval	$F(13, 18162) = 8.0$ ***	$F(13, 18162) = 90.6$ ***	
Score harmonic interval	$F(11, 18164) = 11.8$ ***	$F(11, 18164) = 13.5$ ***	$F(11, 9076) = 34.5$ ***
Score duration	$F(7, 18168) = 13.8$ ***	$F(7, 18168) = 94.5$ ***	
Piece	$F(1, 18174) = 102.7$ ***	$F(1, 18174) = 132.0$ ***	$F(1, 9086) = 121.5$ ***
Vocal part	$F(1, 18174) = 46.8$ ***	$F(1, 18174) = 58.8$ ***	
Age	$F(9, 18166) = 166.0$ ***	$F(9, 18166) = 59.4$ ***	
Musical background	$F(13, 18162) = 177.8$ ***	$F(13, 18162) = 77.6$ ***	

TABLE II. Results of one-way ANOVAs testing each error type grouped by different factors (***p<.001; **p<.01; *p<.05; NS: not significant).

A. Unison vs Duet Singing Condition

To test our first hypothesis, that the unison condition has lower pitch error and interval errors than the duet condition, a one-way ANOVA was conducted. For testing MAPE and MAMIE, we use only the data from dependent singers (those who can hear their partners),

	Condition		**Significance of Difference**
	Unison	**Duet**	
MAPE	0.3518 ± 0.0057	0.4679 ± 0.0076	$F(1, 9086) = 149.38$, $p < .001$
MAMIE	0.2587 ± 0.0039	0.2637 ± 0.0052	$F(1, 9086) = 0.64$, $p = 0.42$
MAHIE	0.3447 ± 0.0060	0.5243 ± 0.0081	$F(1, 2270) = 262.23$, $p < .001$

TABLE III. Results of one-way ANOVA testing the effect of singing condition on accuracy metrics, expressed as mean value ± the 95% confidence interval.

which is one of the singers in the simplex listening condition and both singers in the duplex condition. Harmonic intervals involve both singers, so we only use the data from the duplex condition for MAHIE. Results show a significant effect of singing condition on MAPE and MAHIE, but not for MAMIE (see Table III). Post hoc comparisons using the Tukey HSD test confirmed that MAPE and MAHIE were significantly lower for the unison condition than for the duet condition.

The results confirmed our hypothesis for MAPE and MAHIE, but not for MAMIE. The reason for the higher MAPE in the duet condition (by 12 cents) may be due to the distraction of someone singing a different note, making it more difficult to sing one's own note than when the partner is singing the same note. For harmonic intervals, the duet condition has twelve different score intervals, while the unison condition has only one score interval, the unison interval. The various score intervals are more difficult to sing in tune, resulting in a higher MAHIE (by 38 cents) for the duet condition.

For MAMIE, there is no significant difference between the unison and duet conditions, so we did not find any influence of singing condition on the tuning of melodic intervals. Since melodic intervals are tuned from one's own previous note, the other singer has no direct effect on the target interval, unlike in harmonic intervals, where the tuning is between the singers. The same argument, however, should also apply to pitch error, where a significant difference was observed. The relationship between the three error measures is complex, as any change in a single pitch will alter all measures. Here we see a tendency that when people sing different parts, their relative tuning to each other and absolute tuning to the initial reference suffer, although their local melodic intervals appear no worse. Given an imperfect partner, we suggest that ideal singing would involve a tradeoff between all three error types.

B. Effect of Listening Condition

Hypotheses 2 and 3 predict that the solo listening condition has less pitch error but greater harmonic interval error than the duplex condition. ANOVA tests were conducted to test whether the four listening conditions have an influence on each measure of accuracy. Since the differences between listening conditions depend on whether singers can hear the voice of their partners, we separate the data from the simplex conditions into two cases: dependent singers and independent singers.

The ANOVA results showed that the effects of listening condition on MAPE, MAHIE and MAMIE were all significant: for MAPE, $F(3, 18172) = 52.16$, $p < .001$; for MAMIE, $F(3, 16956) = 38.77$, $p < .001$; and for MAHIE, $F(2, 9085) = 12.76$, $p < .001$. The ANOVA

	Significance of Difference			
	Solo	NS	***	***
		Simp. Indep.	***	***
			Simp. Dep.	***
				Duplex
MAPE	0.32 ± 0.0058	0.33 ± 0.0058	0.38 ± 0.0058	0.41 ± 0.0058

TABLE IV. Results of Tukey HSD test showing the effect of listening condition (solo, simplex independent, simplex dependent, duplex) on MAPE (***p<.001; **p<.01; *p<.05; NS: not significant). The bottom line shows the mean value ± 95% confidence interval for each group.

test tells whether there is an overall difference between groups, but it does not tell which specific groups differed. Post hoc comparisons using the Tukey HSD test were applied to find out which specific groups differed (Tables IV, V and VI).

The results support hypothesis 2, as the MAPE of the solo condition has 9 cents less pitch error than the duplex condition (Table IV). In general, participants have more pitch error when they can hear their partner singing than when they sing independently. This applies not only to the solo and duplex conditions, but also to the simplex conditions; in all cases, independent singers (solo and simplex independent) have significantly less MAPE than dependent singers (simplex dependent and duplex).

We also observed that the MAPE of dependent singers in the simplex condition is better than that in the duplex condition. This difference can be explained by considering that the

		Significance of Difference	
	Solo	***	*
		Simplex	NS
			Duplex
MAHIE	0.45 ± 0.0041	0.39 ± 0.0041	0.41 ± 0.0041

TABLE V. Results of Tukey HSD test showing the effect of listening condition (solo, simplex, duplex) on MAHIE (***p<.001; **p<.01; *p<.05; NS: not significant). The bottom line shows the mean value ± 95% confidence interval for each group.

		Significance of Difference		
	Solo	**	***	***
		Simp. Indep.	***	***
			Simp. Dep.	NS
				Duplex
MAMIE	0.23 ± 0.0098	0.21 ± 0.0098	0.26 ± 0.0098	0.26 ± 0.0098

TABLE VI. Results of Tukey HSD test showing the effect of listening condition (solo, simplex independent, simplex dependent, duplex) on MAMIE (***p<.001; **p<.01; *p<.05; NS: not significant). The bottom line shows the mean value ± 95% confidence interval for each group.

partner of the dependent singer is an independent singer, while the partner of the duplex singer is a dependent singer. We saw above that independent singers have lower MAPE than dependent singers, and accordingly their partners, who hear them, also sing with less pitch error.

The results for hypothesis 3 are shown in Table V. In agreement with the hypothesis, the duplex condition has less harmonic interval error than the solo condition, even though the pitch error and melodic interval error are greater. For MAHIE, there is also a significant difference between solo and simplex conditions ($p < 0.001$) but not between the simplex and duplex conditions ($p > 0.05$).

As shown in Table VI, dependent singers in the simplex and duplex conditions have more MAMIE than independent singers ($p < 0.001$ in all four cases). These results have a similar pattern to those obtained for MAPE. An unexpected significant difference was found between the two independent conditions (where the singer cannot hear her partner). The effect size is small (2 cents), and can be explained as an order effect, as the solo condition preceded the simplex conditions.

C. Correlation of Dependent and Independent Singers' Errors

We then test hypothesis 4, whether there is a linear relationship between the pitch error (PE) of dependent and independent singers in the simplex condition. A linear regression was performed to model the pitch error of the dependent singer e^p_D as a function of the pitch error of the independent singer e^p_I (Figure 3), using the data from the duet condition only. A significant regression equation was found, $e^p_D = 0.02 + 0.91 e^p_I$ ($p < .001$), with

$R^2 = 0.28$. The unison singing condition also exhibited a significant linear relationship, but with a smaller slope than in the duet condition.

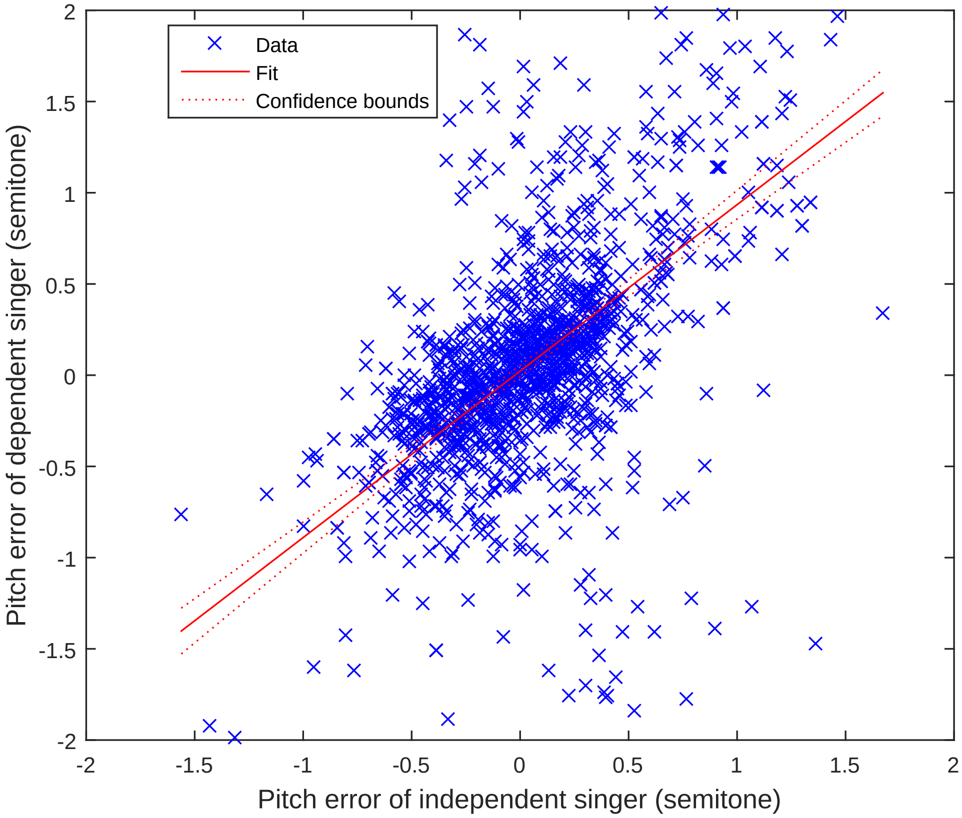


FIG. 3. Scatter plot showing the correlation between independent and dependent singers' pitch error in the duet singing condition and simplex listening condition.

The melodic interval error (MIE) of dependent singers is also positively correlated to the MIE of independent singers ($r = 0.41$, $p < 0.001$) in the duet condition. The weak linear relationship is described by the following formula: $e_D^m = 0.005 + 0.59 e_I^m$, with $R^2 = 0.17$. There was also a significant but weak linear relationship between pitch variation of dependent singers and independent singers ($r = 0.12$, $p < 0.001$).

D. Pitch Variation within Notes

Hypothesis 5 concerns the pitch variation of dependent and independent singers. Pitch variation (Equation 8) does not show any significant effect of listening condition ($\mathsf{F}(3, 17564) = 1.47, \mathsf{p} = 0.22$). Likewise, an ANOVA applied to the two groups dependent singer and independent singer does not show a significant difference ($\mathsf{F}(1, 17566) = 1.74, \mathsf{p} = 0.19$). Thus the results fail to confirm our final hypothesis. We had expected to find evidence of singers adjusting to their partner's pitch during a note. Some pairs of participants show a significant difference, where the pitch variation of dependent singers is higher than that of independent singers, as predicted, but this effect was not consistent across the whole dataset.

Moreover, the pitch variation in the unison condition (mean: 0.09; SD: 0.14) is lower than in the duet condition (mean: 0.11; SD: 0.16), with a statistically significant difference ($\mathsf{F}(1, 17566) = 53.95$, $\mathsf{p} < .001$). The pitch trajectories of the unison condition tend to be flatter in shape than those of the duet condition. There are a few factors that significantly influence pitch variation: the piece ($\mathsf{F}(1, 17566) = 52.61$, $\mathsf{p} < .001$), individual differences ($\mathsf{F}(15, 17552) = 53.62$, $\mathsf{p} < .001$), and score pitch ($\mathsf{F}(15, 17552) = 20.6$, $\mathsf{p} < .001$), where the high pitches (D5, E♭5) in particular exhibit greater variation. Thus pitch variation appears to reflect uncertainty of the singer in trying to reach the intended pitch, rather than deliberate adjustments to improve intonation.

E. Factors Based on the Score

The target pitch and its melodic and harmonic context are also expected to influence singing accuracy. We tested these factors with a series of ANOVAs. Score pitch ($\mathsf{F}(15, 17552) = 22.23$, $\mathsf{p} < .001$), score melodic interval ($\mathsf{F}(13, 18162) = 7.99, \mathsf{p} < .001$) and score harmonic interval ($\mathsf{F}(11, 18164) = 11.8, \mathsf{p} < .001$) all have a significant effect on MAPE. Likewise for MAMIE, score pitch ($\mathsf{F}(15, 16346) = 10.88$, $\mathsf{p} < .001$), score melodic interval ($\mathsf{F}(13, 16946) = 89.02, \mathsf{p} < .001$) and score harmonic interval ($\mathsf{F}(11, 16948) = 13.3, \mathsf{p} < .001$) all have a significant effect.

Although the score pitch has a significant effect on MAPE, the correlation between them does not show a linear trend. It is rather the musical context which dictates which notes elicit larger errors, as shown by the interval-based results below. The most accurate pitch is C4 (0.260 ± 0.009) while the least accurate pitches are A3 (0.514 ± 0.023) and D♯4 (0.452 ± 0.011).

Figure 4 shows the MAMIE for each score interval. The errors group into three clusters corresponding to (absolute) interval size. The unison interval has the smallest error, less than 15 cents, while intervals of one to three semitones have mean errors between 25 and 30 cents, and larger intervals have mean errors between 30 and 45 cents. All differences between clusters are significant, except for the ascending minor 7th (+10 semitone) interval, discussed below, and the ascending major third (+4), which lies on the border between the two clusters. We thus see a general pattern of larger errors for larger intervals, with a small and non-significant tendency for descending intervals to have larger errors than their ascending counterparts. The ascending minor 7th interval is exceptional, being the largest

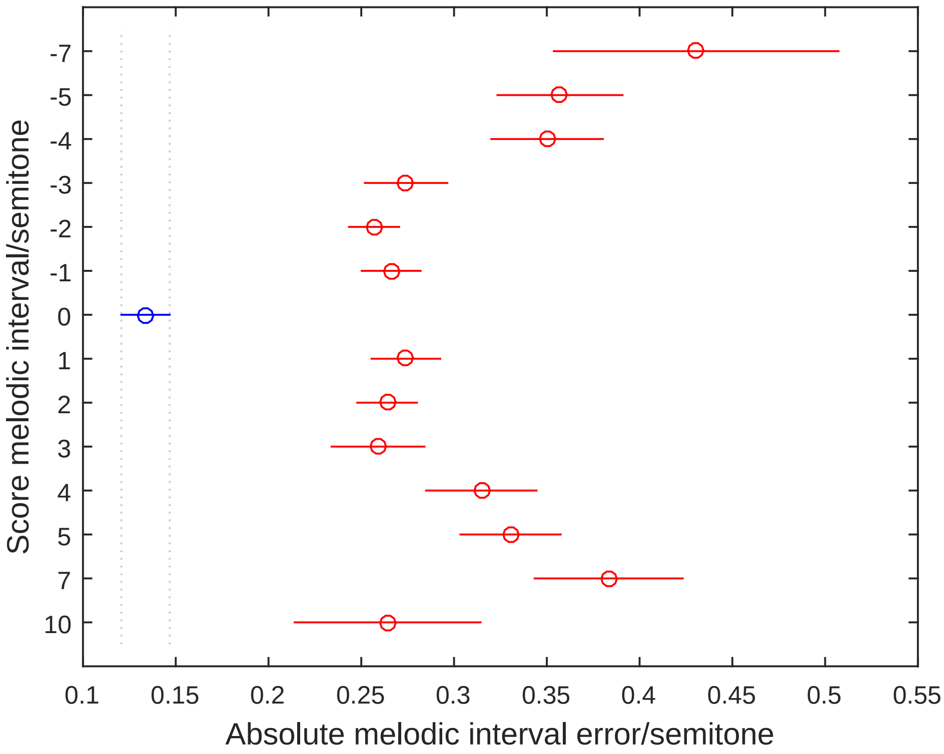


FIG. 4. The mean estimates and the standard errors of absolute melodic interval error for each score melodic interval (significant differences from the unison interval are shown in red).

interval, but having an error in the range of the smaller interval cluster. This interval only occurs twice, both times in the soprano part of the first piece. We believe the lower error is due to the fact that this melody (Silent Night) is particularly well-known.

The score harmonic interval has a significant effect on MAHIE ($\mathsf{F}(11, 9076) = 34.48, \mathsf{p} < .001$), as shown in Figure 5. Again the unison interval has the lowest error, and most score harmonic intervals have significant differences in MAHIE from the unison interval, except the major second and major sixth intervals. The least consonant intervals have the greatest error, with the minor second (mean:0.66; SD=0.98) and diminished fifth (mean:0.67; SD=0.79) having the largest MAHIE and also the largest spread of values.

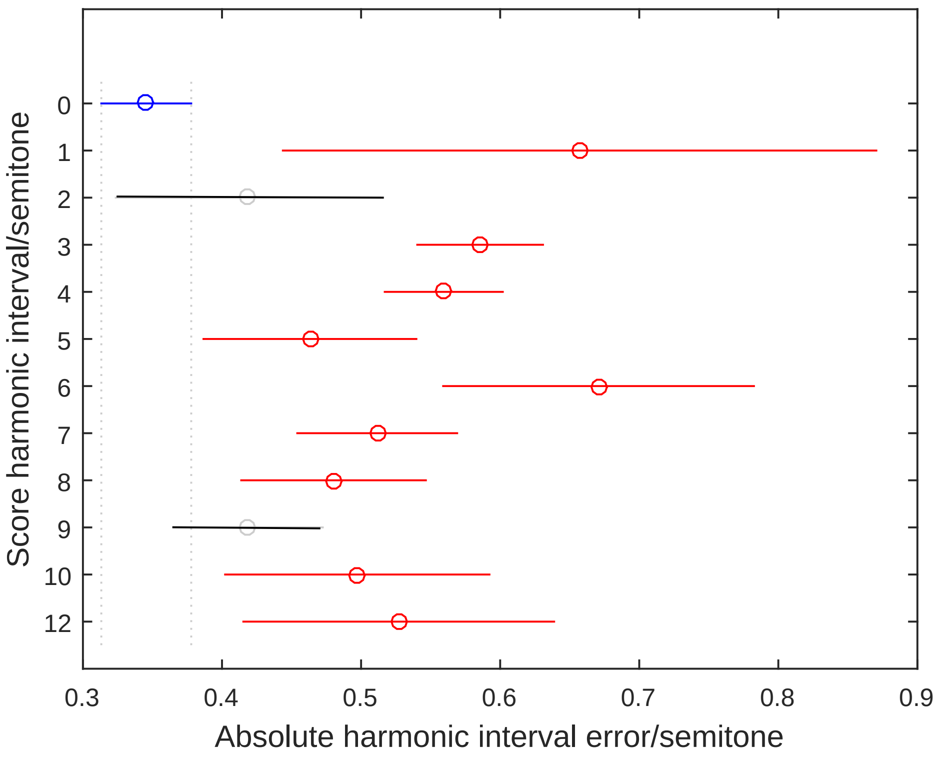


FIG. 5. The mean estimates and the standard errors of absolute harmonic interval error for each score harmonic interval (significant differences from the unison interval are shown in red).

F. Vocal Part

The effect of vocal part (soprano, alto) on intonation accuracy was also investigated. Based on a one-way ANOVA, the vocal part has a statistically significant effect on MAPE ($\mathsf{F}(1, 18174) = 46.78, \mathsf{p} < .001$) and MAMIE ($\mathsf{F}(1, 18174) = 58.76, \mathsf{p} < .001$).

According to Section V A, the unison condition has less MAPE and MAMIE than the duet condition in general. However, we find an interaction with the factor of the vocal part. A two-way ANOVA was performed to examine the effect of singing condition and vocal part on MAPE. There is a significant interaction between the effects of vocal part and singing condition ($F(1, 18172) = 61.96, p < .001$). Simple main effects analysis (Table VII) showed that sopranos have significantly less MAPE than altos in the duet singing condition ($F(1, 6462) = 82.14, p < .001$) but there are no significant differences between vocal parts in the unison condition ($F(1, 11710) = 1.08, p = 0.30$). Further, the MAPE of the soprano part does not change significantly between the unison and duet conditions, but the alto part has a significantly larger MAPE in the duet condition as opposed to the unison condition. For MAMIE in both vocal parts, the duet condition has lower MAMIE than the unison condition, and in both conditions, the alto part has greater MAMIE than soprano.

	Unison	**Duet**	**Significance: singing condition**
MAPE Soprano	0.34	0.34	NS
MAPE Alto	0.34	0.44	***
Significance: vocal part	NS	***	
MAMIE Soprano	0.23	0.21	***
MAMIE Alto	0.26	0.25	**
Significance: vocal part	***	***	

TABLE VII. MAPE and MAMIE of soprano and alto in unison and duet singing conditions, and dependent listening conditions, showing the significance of differences between vocal parts and between singing conditions (***p<.001; **p<.01; *p<.05; NS: not significant).

G. Pitch Drift

Besides the previous factors, the note number in the trial also has a significant influence on MAPE ($F(54, 18121) = 6.44, p < .001$ in Table II). Note number in trial is positively correlated with MAPE, which means that the absolute pitch error increases with time. The regression equation describing the relationship of note number in trial i and MAPE is: $\text{MAPE} = 0.235 + 0.002i$, with $R^2 = 0.016$, $p < .001$. For each adjacent note, MAPE increases by 0.2 cents, resulting in about 10 cents of increase in MAPE from the beginning to the end of each trial.

The direction of the drift varies according to individual differences (Dai *et al.*, 2015; Mauch *et al.*, 2014); there was no overall trend to drift upwards or downwards. The magnitude of drift is similar to that found in a previous study (Mauch *et al.*, 2014), where drift of 13.8 cents over 50 notes was found.

VI. A COMBINED MODEL FOR PITCH ERROR

Section V investigated single factors that influence the pitch accuracy of solo, unison and duet singers. In this section, we fit the investigated factors to a single linear mixed effects model for absolute pitch error, in order to test whether such a joint model can account for the variations in MAPE.

The multiple factors were analysed using linear mixed-effects regression (LMER), using the `fitlme` function in Matlab and MAPE as the dependent variable. LMER has an advantage over standard data aggregation and repeated-measures ANOVA analysis, in that it controls for the variance associated with random factors without data aggregation. Before building the LMER model, the candidate factors were each tested with a one-dimensional linear regression. Some factors such as score pitch, score melodic interval, score harmonic interval, age, musical background and note duration have a significant effect according to the ANOVA test, but their effect is not linear. (Added: Applying simple non-linear transformations to these variables does not change this fact: the effect of pitch and interval depends on the musical context, e.g. the tonality and the consonance or otherwise of the notes (see Figures 4 and 5); age has a limited range; musical background is sparse, dominated by individual factors; and duration is dominated by other score factors (the pitches of the longest and shortest notes).) For the factors which have a linear effect, we add them one by one into the LMER model and compare with the previous model (i.e. without that factor), using 0.05 as the p-value threshold for rejecting insignificant factors.

The resulting model involved singing condition, vocal part, listening condition and note number in trial as fixed effects. As random effects, we have two factors: the individual singer and the piece. Visual inspection of residual plots did not reveal any obvious deviations from normality. P-values were obtained by likelihood ratio tests of the full model with the effect in question against the model without the effect in question. Table VIII shows the resulting LMER model, where all the tested factors are significant. The same process was attempted for MAMIE and MAHIE, but did not give a significant result.

In Section V A, the duplex condition has a larger MAPE than the other listening conditions, but the LMER gives the opposite result. To investigate further, we applied the LMER model to each group of participants individually, and found that the effect size and tendency

Factor	**Coeff.**	**SE**	**Significance**
(Intercept)	0.0014	0.0500	NS
Note number in trial	0.0007	0.0002	**
Unison condition	-0.0378	0.0076	***
Simplex dependent	0.0300	0.0103	**
Simplex independent	0.0235	0.0103	**
Duplex	-0.0459	0.0100	***
Alto part	0.0528	0.0078	***

TABLE VIII. A linear mixed-effects regression model for absolute pitch error, showing coefficient estimate (Coeff.), standard error (SE) and significance level of all predictors in the analysis (***p<.001; **p<.01; *p<.05; NS: not significant).

vary across groups. For 3 of the groups, the duplex condition has a significant positive effect on MAPE, while 4 groups show a significant negative effect size, and one has no significant difference between conditions. (Added: To account for these group differences the model was refitted with random slopes for condition across groups. However, after refitting with random slopes, the listening conditions do not show any significant results in the LMER model.) Other research on individual versus unison singing has similar controversial results. In a pilot study, Smith (1973) observed some fifth and sixth grade children who sang accurately in a group but not alone, and others who sang more accurately alone. Some report a positive effect of unison singing ((e.g. Smith, 1973)) while others report negative results ((e.g. Goetze, 1989)). Our study includes duet as well as unison singing, and we find that listening condition generally has a significant effect on pitch accuracy, but the tendency and effect size vary due to individual differences.

VII. DISCUSSION

It is evident that dependent singers adjusted their pitch influenced by their partners' pitch. An important question to resolve is whether these adjustments were deliberate (e.g. to mitigate inaccuracies in their partner's singing), or inadvertent changes caused by the distraction of the partner's voice. Table V shows that the MAHIE in the simplex and duplex conditions is smaller than in the solo condition ($\mathsf{p} < .001$). At the same time, singers who hear the voice of their partners (dependent singers) have higher MAPE and MAMIE than independent singers. Taken together, this supports the view that singers sacrifice some accuracy in singing their own part in order to harmonise (or sing in unison) better with their partner.

In this work, we report averages across singers (and their partners), not taking into account individual characteristics which may vary from pair to pair, for example the tendency of a singer to lead or follow, regardless of their partner's accuracy. One could characterise such tendencies by the extent of influence of the partner's singing, where a leader would be influenced less and a follower more by their partner's pitch. It is likely that such characteristics of interaction exist and influence the results, but our experimental design (each singer sings with a fixed partner) does not allow us to determine such cases unambiguously, as a singer's behaviour might arise in part from a reaction to their particular partner.

In a standard choral situation, multiple singers are assigned to each of several parts. Our study only considers the simpler case of two singers, and we must use caution in extrapolating to the more general case. Conventionally, conductors group singers with the same vocal part together. The overall lower pitch error for the unison condition supports this practice, although the interaction with vocal part suggests that it might not be necessary for the sake of a dominant part such as soprano. Another choral practice supported by these results is to place weaker singers next to strong singers so that they can intentionally follow their pitch.

Although the participants of this study were selected as having vocal performance and choral experience, they are all amateur singers. They were given limited time to learn their parts (although one can assume that they already knew the melody of Silent Night), so some of the error could be due to lack of familiarity with the parts. We might have obtained different results if we had focused on professional singers, where the overall level of accuracy is likely to have been much higher.

VIII. CONCLUSIONS

This paper presented an experiment investigating pitch accuracy and interaction in unaccompanied duet singing. 16 female participants sang two pieces of music in two singing conditions (unison and duet) and three types of listening condition (solo, simplex and duplex). The results indicated significant effects of the following factors on absolute pitch error: singing condition, listening condition, vocal part, and note number in trial, as well as score factors and individual factors of the singer. Likewise the melodic intervals and the harmonic intervals were affected by the same factors.

In terms of singing conditions, the unison condition has 12 cents less mean absolute pitch error and 38 cents less mean absolute harmonic interval error than the duet condition. This gives some measure of the additional difficulty of singing in harmony, and particularly of tuning non-unison intervals.

The general effect of singing with a partner is an increase in errors of individual pitches and intervals, but a reduction in the error of the interval between singers. That is, singers adjust their pitch to harmonise better with their partner, at the expense of continuity of tonal reference. Independent singers have 7 cents less pitch error than singers who can hear their partner.

The target harmonic interval has a significant effect on MAHIE, with dissonant intervals having the largest errors and the unison interval the smallest. For melodic intervals, the perfect fifth had the largest MAMIE, which is somewhat surprising considering the previous result and the fact that it is a consonant interval. However it is one of the largest melodic intervals in our material (exceeded only by the two minor 7th leaps in the soprano part of Silent Night), and thus we suggest the size of the interval to be a contributing factor in this case. We would expect consonance of intervals to play a smaller role for melodic intervals than harmonic intervals, since the pitches do not sound simultaneously in the melodic case.

We found a positive correlation between the signed pitch errors of dependent singers and independent singers in the simplex condition. In other words, if one singer sings sharp, their partner is influenced to sing sharp as well. The correlation of pitch errors is again evidence of interaction, that singers adjust their pitch to improve harmonic intervals at the expense of melodic intervals and preservation of the tonal reference.

Analysis of the pitch trajectories within tones revealed greater stability of pitch in the unison condition than the duet condition, but not in independent singers over dependent singers. Although stability is correlated with singing accuracy, pitch variation is necessary if singers are to adjust dynamically to the pitch of an imperfect partner, which is what we expected to find in the data. However, our results suggest that the observed pitch variation arises more from imprecision or uncertainty than deliberate adjustment. Further analysis of the pitch trajectories would be an interesting avenue for future work.

We also tested the obtained factors in a combined model using linear mixed-effects regression. The model shows note number in trial, singing condition, listening condition and vocal part have a significant influence on absolute pitch error. More specifically, the absolute pitch error increases about 10 cents over a trial, indicating the existence of pitch drift. The unison condition has 4 cents less absolute pitch error than the duet condition. For singing condition, the simplex conditions involve a small increase in pitch error, in agreement with results in Section V B, but the duplex condition gave a decrease of 5 cents, contrary to the previous results. The effect of the duplex condition varied in direction and size between groups, with some groups performing better together while other groups sing better individually.

There is considerable scope for further work on singing intonation and interaction, either by extending the analysis of the dataset, which is released as open data (Section IX), or by collecting further data for analysis. In particular, in order to move towards more typical musical settings, we would need to investigate cases where there are multiple (more than two) singers per part, multiple parts, and instrumental accompaniment. In a follow-up study, we have recorded several quartets singing in an SATB setting, the preliminary results of which have been reported (Dai and Dixon, 2017).

IX. DATA AVAILABILITY

The code and the data needed to reproduce our results (note annotations, questionnaire results, score information) are available from `https://code.soundsoftware.ac.uk/projects/pitch-accuracy-and-interaction-in-unaccompanied-duet-singing/repository`.

ACKNOWLEDGMENTS

The study was conducted with the approval of the Queen Mary Research Ethics Committee (approval number: QMREC1456).

Many thanks to all of the participants who contributed to this project, including the QMUL *A Capella* Society, QMUL Music Society. We also thank Marcus Pearce and Daniel Stowell for their advice on data analysis. Jiajie Dai is supported by a China Scholarship Council and Queen Mary Joint PhD Scholarship.

Alldahl, P.-G. (**2008**). *Choral Intonation* (Gehrmans, Stockholm, Sweden).

Berkowska, M., and Dalla Bella, S. (**2009**). “Acquired and congenital disorders of sung performance: A review,” Adv. Cogn. Psychol. **5**, 69–83, doi: `10.2478/v10053-008-0068-2`.

Boersma, P. (**2002**). “Praat, a system for doing phonetics by computer,” Glot Int. **5**(9/10), 341–345.

Bohrer, J. C. S. (**2002**). “Intonational strategies in ensemble singing,” Ph.D. thesis, City University, London.

Brandler, B. J., and Peynircioglu, Z. F. (**2015**). “A comparison of the efficacy of individual and collaborative music learning in ensemble rehearsals,” J. Res. Music Educ. **63**(3), 281–297.

Bretos, J., and Sundberg, J. (**2003**). “Vibrato extent and intonation in professional Western lyric singing,” J. Voice **17**, 343–352.

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

Clayton, L. (**1986**). “An investigation of the effect of a simultaneous pitch stimulus on vocal pitch accuracy,” Master’s thesis, Indiana University, Bloomington.

Cooper, N. A. (**1995**). “Children’s singing accuracy as a function of grade level, gender, and individual versus unison singing,” J. Res. Music Educ. **43**(3), 222–231.

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

Dai, J., and Dixon, S. (**2016**). “Analysis of vocal imitations of pitch trajectories,” in *17th Int. Soc. Music Inf. Retr. Conf.*, pp. 87–93.

Dai, J., and Dixon, S. (**2017**). “Analysis of interactive intonation in unaccompanied SATB ensembles,” in *18th Int. Soc. Music Inf. Retr. Conf.*, pp. 599–605.

Dai, J., Mauch, M., and Dixon, S. (**2015**). “Analysis of intonation trajectories in solo singing,” in *16th Int. Soc. Music Inf. Retr. Conf.*, pp. 420–426.

Dalla Bella, S., Giguère, J.-F., and Peretz, I. (**2007**). “Singing proficiency in the general population,” J. Acoust. Soc. Am. **121**(2), 1182–1189, doi: 10.1121/1.2427111.

de Cheveigné, A., and Kawahara, H. (**2002**). “YIN, a fundamental frequency estimator for speech and music,” J. Acoust. Soc. Am. **111**(4), 1917–1930, doi: 10.1121/1.1458024.

Demorest, S. M., Pfordresher, P. Q., Dalla Bella, S., Hutchins, S., Loui, P., Rutkowski, J., and Welch, G. F. (**2015**). “Methodological perspectives on singing accuracy: An introduction to the special issue on singing accuracy (part 2),” Music Percept. **32**(3), 266–271.

Devaney, J., and Ellis, D. P. W. (**2008**). “An empirical approach to studying intonation tendencies in polyphonic vocal performances,” J. Interdiscipl. Music Stud. **2**(1), 141–156.

Ferrante, I. (**2011**). “Vibrato rate and extent in soprano voice: A survey on one century of singing,” J. Acoust. Soc. Am. **130**(3), 1683–1688.

Ganschow, C. M. (**2014**). “Secondary school choral conductors’ self-reported beliefs and behaviors related to fundamental choral elements and rehearsal approaches,” J. Music Teacher Educ. **23**(2), 52–63.

Gerhard, D. (**2003**). “Pitch extraction and fundamental frequency: History and current techniques,” Technical Report TR-CS 2003-06 .

Goetze, M. (**1985**). *Factors Affecting Accuracy in Children’s Singing* (University of Colorado at Boulder).

Goetze, M. (**1989**). “A comparison of the pitch accuracy of group and individual singing in young children,” Bull. Counc. Res. Music Educ. 57–73.

Green, G. A. (**1994**). “Unison versus individual singing and elementary students’ vocal pitch accuracy,” J. Res. Music Educ. **42**(2), 105–114.

Heath, C., and Gonzalez, R. (**1995**). "Interaction with others increases decision confidence but not decision quality: Evidence against information collection views of interactive decision making," Organ. Behav. Hum. Decis. Process. **61**(3), 305–326.

Hess, W. (**1983**). *Pitch Determination of Speech Signals: Algorithms and Devices* (Springer, Berlin, Germany).

Howard, D. M. (**2003**). "A capella SATB quartet in-tune singing: Evidence of intonation shift," in *Stockholm Music Acoust. Conf.*, Vol. 2, pp. 462–466.

Howard, D. M. (**2007**). "Intonation drift in *A Capella* soprano, alto, tenor, bass quartet singing with key modulation," J. Voice **21**(3), 300–315.

Kalin, G. (**2005**). "Formant frequency adjustment in barbershop quartet singing," Master's thesis, KTH Royal Institute of Technology, Stockholm.

Liimola, H. (**2000**). "Some notes on choral singing," in *The Cambridge Companion to Singing*, edited by J. Potter (Cambridge University Press, Cambridge, UK), pp. 149–157, doi: `10.1017/CCOL9780521622257.013`.

Lim, M. C. (**2014**). "In pursuit of harmony: The social and organisational factors in a professional vocal ensemble," Psychol. Music **42**(3), 307–324.

Lindley, M. (**2001**). "Just intonation," Grove Music Online, edited by L. Macy. http://www.grovemusic.com (accessed 30 January 2015) .

Long, P. A. (**1977**). "Relationships between pitch memory in short melodies and selected factors," J. Res. Music Educ. **25**(4), 272–282.

Mauch, M., Cannam, C., Bittner, R., Fazekas, G., Salamon, J., Dai, J., Bello, J., and Dixon, S. (**2015**). "Computer-aided melody note transcription using the Tony software: Accuracy and efficiency," in *Proceedings of the First International Conference on Technologies for Music Notation and Representation*, pp. 23–30.

Mauch, M., and Dixon, S. (**2014**). "PYIN: A fundamental frequency estimator using probabilistic threshold distributions," in *IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 659–663.

Mauch, M., Frieler, K., and Dixon, S. (**2014**). "Intonation in unaccompanied singing: Accuracy, drift, and a model of reference pitch memory," J. Acoust. Soc. Am. **136**(1), 401–411.

Morrell, M. J., Harte, C. A., and Reiss, J. D. (**2011**). "Queen Mary's 'Media and Arts Technology Studios' audio system design," in *Audio Engineering Society Convention 130*, Audio Engineering Society.

Mullen, P. (**2000**). "Pitch drift as a result of just intonation," J. Acoust. Soc. Am. **108**(5), 2618.

Müllensiefen, D., Gingras, B., Musil, J., and Stewart, L. (**2014**). "The musicality of non-musicians: An index for assessing musical sophistication in the general population," PLoS ONE **9**(2), e89642.

Mürbe, D., Pabst, F., Hofmann, G., and Sundberg, J. (**2002**). "Significance of auditory and kinesthetic feedback to singers' pitch control," J. Voice **16**(1), 44–51.

Nichols, B. E. (**2016**). "Task-based variability in children's singing accuracy," J. Res. Music Educ. **64**(3), 309–321.

Nichols, B. E., and Wang, S. (**2016**). "The effect of repeated attempts and test-retest reliability in children's singing accuracy," Music. Sci. **20**(4), 551–562.

Pfordresher, P. Q., and Brown, S. (**2007**). "Poor-pitch singing in the absence of 'tone deafness'," Music Percept. **25**(2), 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**(4), 2182–2190.

Pfordresher, P. Q., and Mantell, J. T. (**2014**). "Singing with yourself: Evidence for an inverse modeling account of poor-pitch singing," Cogn. Psychol. **70**, 31–57.

Potter, J. (**2000**). "Ensemble singing," in *The Cambridge Companion to Singing*, edited by J. Potter (Cambridge University Press, Cambridge, UK), pp. 158–164, doi: `10.1017/CCOL9780521622257.014`.

Prout, E. (**2011**). *Harmony: Its Theory and Practice* (Cambridge University Press).

Rabiner, L., Cheng, M., Rosenberg, A., and McGonegal, C. (**1976**). "A comparative performance study of several pitch detection algorithms," IEEE Trans. Acoust., Speech, Signal Process. **24**(5), 399–418.

Seashore, C. E. (**1914**). "The tonoscope," Psychol. Mono. **16**(3), 1–12.

Simpson, J. A., Weiner, E. S. *et al.* (**1989**). *The Oxford English Dictionary*, **2** (Clarendon Press, Oxford).

Smith, R. S. (**1973**). "Factors related to children's in-tune singing abilities," Ph.D. thesis, West Virginia University.

Sundberg, J. (**1987**). *The Science of the Singing Voice* (Northern Illinois University Press, DeKalb, IL).

Takeuchi, A. H., and Hulse, S. H. (**1993**). "Absolute pitch," Psychol. Bull. **113**(2), 345–361.

Terasawa, H. (**2004**). “Pitch drift in choral music” Music 221A final paper, Center for Computer Research in Music and Acoustics, Stanford University, URL https://ccrma.stanford.edu/~hiroko/pitchdrift/paper221A.pdf.

Vurma, A., and Ross, J. (**2006**). “Production and perception of musical intervals,” Music Percept. **23**(4), 331–344.

Warren, R. A., and Curtis, M. E. (**2015**). “The actual vs. predicted effects of intonation accuracy on vocal performance quality,” Music Percept. **33**(2), 135–146.

Watts, C., Barnes-Burroughs, K., Andrianopoulos, M., and Carr, M. (**2003**). “Potential factors related to untrained singing talent: A survey of singing pedagogues,” J. Voice **17**(3), 298–307.

Welch, G. F., Sergeant, D. C., and White, P. J. (**1997**). “Age, sex, and vocal task as factors in singing ‘in tune’ during the first years of schooling,” Bull. Counc. Res. Music Educ. **133**, 153–160.

Zatorre, R. J., and Baum, S. R. (**2012**). “Musical melody and speech intonation: Singing a different tune,” PLoS Biol. **10**(7), e1001372, http://dx.plos.org/10.1371/journal.pbio.1001372, doi: 10.1371/journal.pbio.1001372.

List of Changes

Added: Applying simple non-linear transformations to these variables does not change this fact: the effect of pitch and interval depends on the musical context, e.g. the tonality and the consonance or otherwise of the notes (see Figures 4 and 5); age has a limited range; musical background is sparse, dominated by individual factors; and duration is dominated by other score factors (the pitches of the longest and shortest notes). , on page 35, line 522.

Added: To account for these group differences the model was refitted with random slopes for condition across groups. However, after refitting with random slopes, the listening conditions do not show any significant results in the LMER model., on page 36, line 542.