Singing ability is rooted in vocal-motor control of pitch

Sean Hutchins · Pauline Larrouy-Maestri · Isabelle Peretz

Abstract The inability to vocally match a pitch can be caused by poor pitch perception or by poor vocal-motor control. Although previous studies have tried to examine the relationship between pitch perception and vocal production, they have failed to control for the timbre of the target to be matched. In the present study, we compare pitch-matching accuracy with an unfamiliar instrument (the slider) and with the voice, designed such that the slider plays back recordings of the participant's own voice. We also measured pitch accuracy in singing a familiar melody ("Happy Birthday") to assess the relationship between single-pitch-matching tasks and melodic singing. Our results showed that participants (all nonmusicians) were significantly better at matching recordings of their own voices with the slider than with their voice, indicating that vocal-motor control is an important limiting factor on singing ability. We also found significant correlations between the ability to sing a melody in tune and vocal pitch matching, but not pitch matching on the slider. Better melodic singers also tended to have higher quality voices (as measured by acoustic variables). These results provide important evidence about the role of vocal-motor control in poor singing ability and demonstrate that single-pitch-matching tasks can be useful in measuring general singing abilities.

Introduction

Music is an important and universal aspect of culture, and one of the most prevalent forms of musical activity is singing. Nevertheless, many people do not sing well, despite having no problems hearing or understanding music. Poor singing ability can have several manifestations, including problems with timing and with timbre (the quality of a sound, independent of pitch, loudness, or timing). However, the most common manifestation of poor singing ability is poor pitch control (e.g., Dalla Bella, Giguère, & Peretz, 2007). In addition, music educators rank pitch intonation as the single most important factor in determining someone's singing talent (Watts, Barnes-Burroughs, Andrianpoulos, & Carr, 2003). Because of this, many studies of singing ability have focused on the ability to match one or more pitches (e.g., Estis, Coblentz, & Moore, 2009; Hutchins & Peretz, 2012; Pfordresher & Brown, 2007; Watts, Moore, & McCaghren, 2005). This ability is commonly measured acoustically as the distance in pitch between the sung note and the target note (the error).

In order to vocally match a target note, singers must perceive the pitch, determine the configuration of their vocal apparatus that will create a note of the same pitch, and enact that motor command. An error in any of these steps will lead to inaccurate vocal pitch matching (Berkowska & Dalla Bella, 2009; Pfordresher & Brown, 2007). Many studies searching for the cause of poor singing abilities have focused on perceptual abilities, generally by correlating vocal pitch-matching accuracy with a measure of pitch perception. Although some of these studies have found relationships between these two abilities (e.g., Estis et al., 2009; Estis, Dean-Claytor, Moore, & Rowell, 2011; Moore, Keaton, & Watts, 2007; Watts et al., 2005), several others have failed to do so (e.g., Bradshaw & McHenry, 2005; Dalla Bella et al., 2007; Moore, Estis,

S. Hutchins (✉)
The Royal Conservatory of Music, 273 Bloor St. West, Toronto, Ontario, Canada M5S 1V6
e-mail: sean.hutchins@gmail.com

P. Larrouy-Maestri
Department of Psychology: Cognition and Behaviour, University of Liège, Liège, Belgium

P. Larrouy-Maestri
Department of Psychology, University at Buffalo, State University of New York, Oswego, NY, USA

S. Hutchins · P. Larrouy-Maestri · I. Peretz
International Laboratory for Brain, Music and Sound Research (BRAMS) and Center for Research on Brain, Language and Music (CRBLM), Université de Montréal, Montréal, QC, Canada

Gordon-Hickey, & Watts, 2008; Pfordresher & Brown, 2007). Ultimately, however, it is unlikely that problems with pitch perception can account for the majority of instances of poor pitch-matching abilities, given that errors in vocal pitch matching (when they occur) are often much larger than errors in measured pitch perception ability.

A recent study by Hutchins and Peretz (2012) used a novel method to investigate the relationship between the ability to perceive and vocally imitate pitches. This study used a new instrument called a slider as a nonvocal alternative to pitch matching. The slider is played by pressing on a horizontal touch-sensitive strip and creates a synthetic vocal tone based on the position at which it is pressed. Unlike a piano, this instrument is not divided into discrete steps but can create any pitch within its range, just like the voice can. Hutchins and Peretz asked musicians and nonmusicians to match synthesized vocal tones on the slider and with their own voices. Participants were considerably better at matching pitches with the slider than with their voices, despite their unfamiliarity with the new instrument. As a control for timbre, participants were also asked to match examples of their own voice singing, with the target examples taken from prior recordings. Participants were more accurate at matching their own voices than synthesized tones but were still less accurate than they were with the slider. Overall, the pattern of results indicated that poor vocal pitch-matching ability was not generally caused by poor perception ability, since poor singers were generally able to match pitches accurately on the slider. Rather, of the 31 nonmusicians tested, 20% were impaired on both singing tasks (matching synthesized vocal tones and matching their own voice), indicating a vocal-motor impairment, and a further 35% were impaired at matching synthesized vocal tones, but not at matching their own voice, indicating a sensorimotor problem involving translating between timbres (since they could vocally match only an identical timbre, but not a different timbre; see Hutchins & Peretz, 2012, for further details). All participants, however, were more accurate in using the slider than in either kind of singing.

An alternative explanation for the general superiority on the slider than on vocal self-matching is related to the specific timbres used in each task. Participants were able to make more accurate tuning judgments (deciding whether two notes were the same or one was mistuned) about the synthesized vocal timbre created by the slider than about natural vocal timbres (Hutchins & Peretz, 2012, Experiment 5). This general trend to be less discerning of tuning errors in the voice than in other instruments, termed the *vocal generosity effect*, was confirmed and extended in a later study (Hutchins, Roquet, & Peretz, 2012). Most listeners do not notice tuning errors in the voice until they reach 50 cents off (100 cents = 1 semitone), as compared with only 20–30 cents in an instrument (such as the slider). Thus, participants' lower errors on the slider task than on the self-matching task may only have been a function of a better ability to resolve tuning for the slider's timbre.

Another question that can be raised about the study of Hutchins and Peretz (2012), as well as many other studies of singing ability, is the relationship between the ability to sing single tones and whole melodies. Many studies of singing ability focus on pitch matching of single tones or small numbers of tones, with the assumption that this ability will scale up to the ability to accurately sing whole melodies (with a few exceptions; e.g., Dalla Bella et al., 2007). However, this assumption has not yet been validated experimentally. Pfordresher and Brown (2007) did show that error on single-pitch matching is correlated with pitch errors in two- or four-tone contexts. However, there are many reasons why single-pitch matching might not always predict melodic singing ability, such as the beneficial effect of establishing a tonal context or the cumulative effect of corrective errors (a negative lag 1 correlation). Given that the ability to sing melodies accurately is generally of much greater importance than matching individual pitches when evaluating overall singing ability, it is important to quantify this relationship.

Finally, one other issue concerns the relationship between singing ability and other measures of vocal quality. Although the mean pitch error is the primary factor in most evaluations of singing ability (Watts et al., 2003), other factors can play a role in how we evaluate singing ability. Our intuition is that poor pitch singers would also tend to have lower quality vocal timbres, but it is possible that these abilities are unrelated, which would mean that there exists a category of singing problems that remains unevaluated by pitch-error-based measurements. In addition, the vocal generosity effect (Hutchins et al., 2012) also demonstrates that timbre can have a significant effect on our tuning judgments (see also Larrouy-Maestri, Magis, & Morsomme, in press), making it important to evaluate these qualities separately. Higher quality vocal tones may assist in evaluating one's own vocal feedback.

To address these questions, we designed a new experiment to compare pitch-matching ability on the slider and voice with each other and with melodic singing ability, without the complications of a varying timbre. Nonmusicians (chosen here because they show a greater range of singing abilities and are more representative of the general population) first recorded single tones at different pitch heights to use as target stimuli. They imitated those recordings with their voice and on the slider. In contrast with the previous studies (Hutchins & Peretz, 2012), the slider sound was not synthesized but used the recordings of each participant's own voice. These recordings were altered using a digital signal processor specifically designed for real-time vocal manipulations. Thus, the vocal pitch-matching responses and the slider-based pitch-matching responses had the same timbre (although self-produced responses will sound somewhat different from slider-produced responses, due to the influence of bone conduction and other physical intermediaries). Following the pitch-matching tasks, participants were also asked to sing a complete version of a well-known song ("Happy

Birthday"). All sung tones were analyzed acoustically for pitch, and the recorded target tones were also analyzed for vocal quality (pitch stability, jitter, and shimmer).

If timbral differences were responsible for participants' better performance on the slider than on the singing tasks in Hutchins and Peretz (2012), then here we should see no difference in pitch error between singing and slider pitch-matching tasks. However, we hypothesize that we will continue to observe better performance in the slider condition than in the singing condition, even when the timbres are the same in both tasks, which would provide confirmatory evidence that singing problems are caused by sensorimotor and vocal-motor impairments, rather than perceptual impairments. We also hypothesize that vocal pitch-matching ability, but not slider pitch-matching ability, should correlate with melodic singing ability, which would confirm the greater importance of vocal motor abilities than perception abilities in determining overall singing ability. Finally, we expect to see that participants with better measurements of vocal quality will be better at vocal pitch matching, since those who have had more practice singing would tend to improve in both vocal pitch and quality and a better tone quality may improve the singer's auditory feedback.

Method

Participants

The participants were 22 nonmusicians (16 female), recruited from university students in Montréal. All participants had 1 year or less of formal training ($M = 0.2$ years) and ranged in age from 18 to 30 years ($M = 23$ years). They reported a mean of 0.32 years of informal musical experience, a mean of 0.32

years of group singing experience, and no formal singing training. No subjects reported any diagnosed hearing deficits or neurological disorders.

Stimuli and procedure

The session was divided into four sections: the recording of target tones from the participants' own voice, a slider (instrument-based) pitch-matching task, a vocal pitch-matching task, and a singing by memory task, and the total experiment lasted approximately 1 h. All vocal stimuli and responses were recorded with a Neumann TLM 103 microphone (Georg Neumann GmbH, Berlin, Germany). Video examples of target recording and the two pitch-matching tasks can be viewed at http://www.brams.umontreal.ca/slidervoice.

Recording targets

In the first section of the experiment, participants recorded their own voice at five different pitch levels, all on the syllable /ba/, sustained for 2–3 s. They sang a low tone, a medium-low tone, a medium tone, a medium-high tone, and a high tone. Tones were self-selected, to ensure that each was within the participant's comfortable range. Three different versions of each pitch level were recorded through Max/MSP (Cycling '74, San Francisco, CA). After all tones had been recorded, the experimenter chose the best example from each of the five categories, using the criteria of pitch stability, voice quality, and differentiability from other pitches. These tones were then normalized for amplitude and trimmed to remove silence from the beginning and end. The five targets were used as the five target tones for the pitch-matching tasks. Figure 1 shows the range of targets produced across all participants.
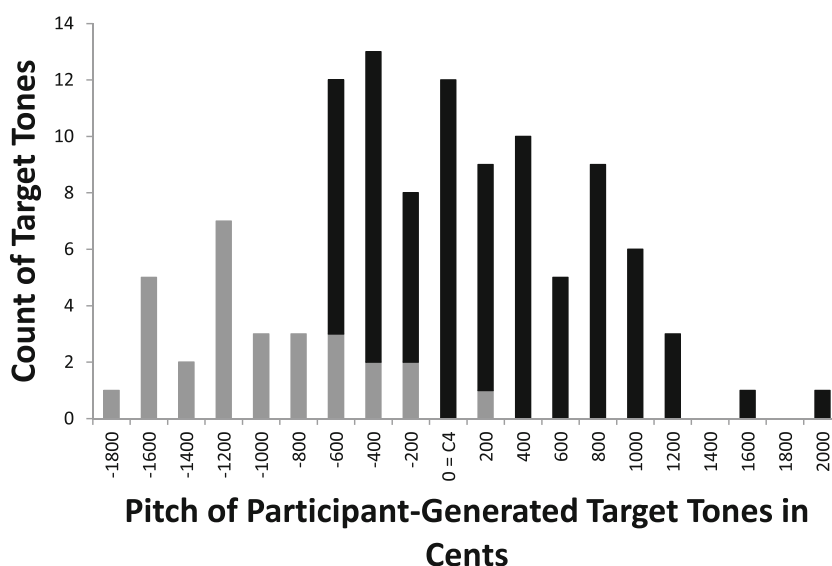
Fig. 1 Histogram of the target tones chosen, sorted by frequency, in bins of 200 cents. Middle C (C4, 261 Hz) is represented as 0. The target tones of males are shown in gray; those of females in black

Slider pitch matching

In the slider task, participants were presented with one of their previously recorded sung tones as a target, and their task was to use the slider to find the same pitch as the target tone. The slider, designed by Hutchins and Peretz (2012), provides a nonvocal pitch-matching measurement that can be compared with vocal pitch matching. The slider produces a pitch based on the position of a finger press and is designed to be easily used by nonmusicians. The slider is made from a 50-cm position sensor overlaid on a pressure sensor (Infusion Systems, Montreal, Canada). The slider can register 1,024 unique positions, making each position less than half a millimeter apart.

On each trial, one of these positions was randomly chosen as the target position (not including the top or bottom one sixth of the slider—170 positions on each side—to ensure adequate space for adjustment in either direction). Any press on the slider triggered the playback of the recording of the target vocal tone. This playback was routed through a VoicePro digital signal processor (TC-Helicon, Victoria, BC, Canada), programmed to apply a pitch shift to the recording of the target vocal tone. The applied pitch shift could range anywhere from plus 10 semitones to minus 10 semitones, and the specific shift was determined by the distance between the target position and the position of the finger press.[1] Finger presses on the slider to the right of the target position triggered a positive (sharp) shift, and those to the left triggered a negative (flat) shift, giving the slider the same pitch orientation as a piano. Each discrete position on the slider represented a pitch shift of 1.17 cents, chosen so that the pitch distance between the two extremes of the slider was equivalent to one octave (1,200 cents = 12 semitones = 1 octave). For example, pressing on the slider 4.17 cm to the right of the target position would yield a difference of +85 steps from the target and would trigger the output of the vocal recording shifted upward by 1 semitone. The shifted output is identical to the original in duration, amplitude, and internal pitch change (e.g., instability, vibrato)

but is shifted globally in pitch. Pressing precisely on the target position plays the original target recording, unshifted.

Each trial was initiated by the participant pressing the space bar and began with an automatic presentation of the target tone, without any shift applied to it. Following this initial presentation, the participant pressed on the slider and heard the resulting shifted version of the target. The original target (without any shift) was played immediately after this, giving participants the ability to compare the pitch of subsequent versions (minimizing the role of pitch memory). Participants could also press the Enter key at any time to relisten to the original target (although they rarely chose to). The target playback was halted whenever the slider was pressed, in order to avoid the superposition of the target sound and the slider-generated sound. These factors make the design very similar to that of Hutchins and Peretz (2012).

Participants were unaware of the randomly chosen target position on each trial and were instructed to find the position on the slider that generated a sound as close as possible to the target tone. They were told that they could respond as many times as they wished until they found the best pitch match for the target and that their accuracy would be judged only by their final response. This helped to ensure that their accuracy on each trial reflected their abilities to perceive the relationship between their produced tone and the target tone, rather than being overly influenced by any unfamiliarity with the instrument itself. Unlike with the voice, motor control of the arm–hand–finger system is known to be quite accurate (De Nil & Lafaille, 2002; Fitts, 1954); thus, we can be confident that response accuracy on the slider is primarily driven by perception ability. Participants indicated that they had finished matching the target by pressing on the space bar to end the trial.

Participants were familiarized with the slider before the task and were allowed to use it freely to become acquainted with it. After this familiarization, they were presented with 5 practice trials. The main experiment consisted of 75 trials, with 15 instances of each of the five target tones, presented in pseudorandom order. Due to time limitations, not all participants could complete all the trials, but each participant completed a minimum of 25 trials ($M = 59.14$ trials, $SD = 17.71$; we have observed in prior experiments that this is approximately the minimum number of trials necessary to obtain sufficiently precise results about a participant's overall accuracy). All data from each trial, including the position of the target, the time and position of each press on the slider, and the shift applied to each output, were recorded and saved as .txt files through Max/MSP.

Vocal pitch matching

The vocal pitch-matching task used the same trial design as the slider pitch-matching task but differed only in the medium of

---

[1] It is a possibility that the large pitch shifts at the extremes of the slider may have introduced distortions that could have provided a timbral cue to target finding. This is a problem for any method of shifting a nonsynthesized pitch. However, we believe that this would have minimal effect on the results, for three reasons. First, the digital signal processor (VoicePro) used to create the pitch-shifted versions of the target is designed for the voice, and other testing in our lab has shown very low levels of audible distortion for pitch shifts up to 2 semitones. Second, even unshifted versions of the tone are routed through the signal processor, with an applied shift of 0 cents. Finally, the effects of the signal distortion would decrease as the participant moved closer to the target position on the slider, making timbral cues less useful for fine adjustments. Because participants were in general very accurate on the slider (13 cents error on average, with 98% of responses within 50 cents), it seems unlikely that timbral cues would be the primary source of feedback for determining the final response.

the response. The order of these two tasks was counterbalanced across participants. The vocal pitch-matching section was preceded by 5 practice trials. This section consisted of 100 trials, with 20 instances of each of the five target tones (this is due to the greater variability in vocal pitch matching and the shorter average time to complete each trial in this condition; more trials are necessary to reach an equivalent estimate of accuracy than with the slider; see Hutchins & Peretz, 2012), presented in pseudorandom order. Each participant completed a minimum of 77 trials ($M = 97.23$ trials, $SD = 4.74$).

As in the slider pitch-matching task, each trial was initiated by the participant pressing the space bar and began with a presentation of the target tone. Participants were instructed to match the target tone as closely as possible using their voice and were told that they could make as many attempts as they liked to match the target. They were asked to wait until the self-recorded target tones had finished playing before singing and could press on the Enter key to relisten to the target.[2] Participants were instructed that their accuracy would be judged only by their final response. Participants indicated that they had finished matching the target by pressing on the space bar to end the trial. The participants' voices were recorded as .aif files.

Melodic singing

In the final section of this experiment, participants were asked to perform the song "Happy Birthday" and were not given a particular starting note. The instruction was to sing "naturally, whilst imagining a festive and friendly context." Participants were informed about the aim of this task—that is, the observation of the pitch accuracy in a melodic context. The participants' sung performances were saved as .wav files.

Analyses

Vocal quality analysis

Acoustic measurements of voice quality were made for each target tone performed by participants in the first section of the experiment (the tones presented in Fig. 1) using Praat (Boersma & Weenink, 2013). We measured the standard deviation of the fundamental frequency (F_0SD) across the duration of the note, as an indication of its pitch stability, the jitter (a measurement of the period perturbation, in

percentages), and the shimmer (a measurement of the amplitude perturbation, in percentages). The latter two are commonly used for the evaluation of voice disorders (Sataloff, 2005; Titze, 2000). Note that for those three measurements, a high score shows a high perturbation of the auditory signal and is associated with a lower voice quality.

Pitch-matching analysis

The vocal recordings (both targets and matching responses) were analyzed using a MATLAB (The MathWorks Inc., Natick, MA) implementation of YIN (de Cheveigné & Kawahara, 2002). These analyses provided information about frequency, amplitude, and aperiodicity at a rate of 1378 Hz, and the pitch information was converted to cents relative to the target pitch. As per the instructions to the participant, only the final response on each trial was evaluated for pitch, but we did compile the number of responses the participant chose to make on each trial.

Pitch-matching accuracy was measured in three ways. First, we measured the signed pitch error, which measured how sharp or flat the response was. Second, we took the absolute value of the pitch error, to avoid sharp and flat errors from canceling each other out in averaging (but see Pfordresher & Mantell, 2009, for a useful alternative to this). Finally, we measured accuracy by taking the percentage of final responses within 50 cents of the target. This criterion was chosen because it is the point at which nonmusicians tend to notice tuning inaccuracies when comparing two vocal tones (Hutchins & Peretz, 2012; Hutchins et al., 2012).

Melodic singing analysis

The song "Happy Birthday" is composed of 25 notes, with one syllable per note. Data processing was done in two stages, using AudioSculpt and OpenMusic (Ircam, Paris, France). The acoustical analyses were based on the extraction of the $F0$ from the stable part of each note (as determined by the experimenter); the onsets and offsets of each note were omitted. Immediately repeated pitches (four in total; each "-py" of "Happy" in this song) were not considered in this analysis, since they were typically very short in duration.

We measured pitch accuracy in "Happy Birthday" in two ways, on the basis of the methods presented in Larrouy-Maestri and Morsomme (2014). First, we calculated the mean interval error as the mean difference between the target interval and the produced interval (unsigned). For example, the first interval, between "Happy" and "Birth-," should be two semitones, 200 cents. If the participant actually sang an interval of 250 cents, that would result in an interval error of 50 cents. Second, we calculated the tonal drift in the same way but ignored all but the most tonally stable notes. In this case, only the first note of the first three phrases (i.e., the first

---

[2] A pilot version of this study played back the unaltered target recording immediately following each vocal response, to match the design of the slider task. However, participants almost invariably chose to respond only once and ended the trial without relistening to the original target (this was also the case in Hutchins & Peretz, 2012). The extra presentation proved to be more confusing than helpful, and our previous results indicate that participants do not become more accurate over multiple attempts to match the same target (Hutchins & Peretz, 2012, Experiment 4). Thus, we dropped the extra presentation and included an option for the participants to listen to the target again if they desired.

syllable of "happy") and the final note of the song were analyzed (the three target intervals were adjusted accordingly, yielding 0, 0, and +500 cents). This provides a measure of how accurately the singers returned to the tonally stable notes, with higher scores indicating that the singer had a greater tendency to allow these to drift over the course of the melody.

Results

Slider–voice comparisons

Participants were considered to be poor at a pitch-matching task if they had a mean absolute pitch error of greater than 50 cents within one response modality (for more on the use of this criterion, see Hutchins & Peretz, 2012; Hutchins et al., 2012). Five participants (23%) passed this threshold in the vocal pitch-matching task, but only 1 (5%) did so in the slider pitch-matching task (who was not one of the 5 who failed the vocal task). Comparisons were carried out between the results from slider and voice pitch-matching tasks for each of the three pitch measurements using three separate paired t tests. We also tested the correlations between slider and voice for each measurement.

On average, participants had lower absolute pitch errors and higher percentage of accurate answers in the slider condition (mean error = 13 cents, $SE = 0.29$ cents, 98% accuracy) than in the vocal condition (mean error = 38 cents, $SE = 1.03$ cents, 86% accuracy), $t(21) = 2.54$, $p = .02$, $d = 0.54$ (error measurement); $t(21) = 2.86$, $p = .009$, $d = 0.62$ (percent accuracy measurement). Participants were able to match a pitch more accurately on the slider than with the voice. There was no significant correlation between slider and voice pitch-matching abilities with either measurement.

The signed pitch error, however, showed a different pattern of results. Here, there was no difference in the tendency to produce flat or sharp errors between slider ($M = 2.01$ cents, $SE = 0.24$) and voice ($M = 0.41$ cents, $SE = 0.92$) conditions, $t(21) = 0.19$, n.s. One-sample t tests indicated that neither modality had a significant tendency to elicit sharp or flat errors, $t(21) = 1.10$, n.s. (slider); $t(21) = 0.05$, n.s. (voice). However, with the signed errors, there was a moderate but significant correlation between the slider and the voice modalities with the signed pitch error measurement, $r(20) = .45$, $p = .04$, indicating that participants who showed the tendency to be more sharp on the slider also tended to be more sharp in their sung responses. However, this was entirely driven by one outlier, and the correlation drops to $r(19) = .15$, n.s., when it is removed.

We also compared the average number of responses that participants chose to make using the slider and their voice during each trial. As in Hutchins and Peretz (2012), participants made more responses in the slider condition ($M = 10.73$, $SE = 0.15$) than in the singing condition ($M = 1.05$, $SE = 0.01$),

$t(21) = 8.20$, $p = .002$, $d = 1.74$. Despite being allowed to correct their responses freely, participants generally chose to respond only once in the singing condition, even though they were less accurate in this condition. It is unlikely that attempting to vocally match the pitch more often would improve their mean error, however, since previous work has shown that participants do not change their pitch-matching accuracy over multiple attempts at matching a single target tone, even over as many as 20 attempts (Hutchins & Peretz, 2012, Experiment 4). The trial-by-trial data in this experiment show the same pattern; our results showed that the average error on the slider did not differ between the first 10 trials ($M = 14.47$) and the 15th through 25th trials (the last 10 completed by each participant; $M = 13.97$), nor did average error differ between the first and last 10 attempts in the vocal-matching condition ($M = 36.80$ vs. $M = 35.98$). More attempts do not lead to greater vocal accuracy.

Vocal quality

To assess how the quality of the original target tones sung by the participants varied according to pitch height, we conducted three separate one-way repeated measures ANOVAs using the factor of target height (five levels), with the measurements of jitter, shimmer, and pitch stability as dependent variables (shown in Table 1; degrees of freedom are reported using Greenhouse–Geisser corrections). We found main effects of jitter, $F(2.67, 55.93) = 18.58$, $p < .001$, $\eta_p^2 = .47$, and shimmer, $F(2.44, 48.95) = 10.61$, $p < .001$, $\eta_p^2 = .37$, such that both tended to decrease (indicating better voice quality) with higher pitches. There was no main effect of target height of the tone on pitch stability. There was also a significant correlation between jitter and shimmer across the five target tones, $r(20) = .49$, $p = .02$, but no correlations between pitch stability and either of the other two measurements. Participants tended to have better voice quality when they sang higher pitches, but this did not affect pitch stability.

Target pitch height

Another relevant question is whether high versus low tones tend to induce errors in particular directions, both with the

Table 1 Means of each vocal quality measurement (shimmer, jitter, and pitch stability [F_0SD]), across the five target pitch heights

	Target Pitch Height				
	Low	Medium-Low	Medium	Medium-High	High
F_0SD	9.01	10.13	12.40	8.29	11.16
Jitter	0.72	0.56	0.46	0.37	0.33
Shimmer	4.44	3.93	3.16	2.59	2.24

slider and with the voice. To measure this, we conducted two one-way repeated measures ANOVAs separately for slider and voice response modalities, across the five levels of target height. The target height category (low, mid-low, medium, mid-high, and high) was used as the independent variable in this analysis, to allow comparison across participants with different vocal ranges (otherwise, this analysis would be dominated by the difference between males and females; see Fig. 1 for the range of actual targets produced). Signed pitch error was used as the dependent variable, to preserve error directionality (see Fig. 2). This measurement was different among targets in the voice condition, $F(1.50, 31.45) = 4.60$, $p = .03$, $\eta_p^2 = .18$. Participants were more likely to sing flat when matching high target notes and more likely to sing sharp when matching low target notes, tending to err toward the middle of their vocal range. However, there was no difference in the ability to match the different target tones in the slider condition, $F(2.46, 51.69) = 0.86$, n.s., which suggests that the vocal errors were due to vocal constraints, rather than perceptual constraints.

Melodic singing ability and its relationship to other tasks

Mean error was not significantly different between the melodic condition (mean interval error, $M = 45$ cents, $SE = 4.92$) and vocal pitch-matching condition (absolute pitch error, $M = 38$ cents, $SE = 1.03$). These two measurements were correlated, $r(20) = .48$, $p = .02$ (see Fig. 3). This correlation dropped slightly below the significance threshold when the four overall least accurate singers were removed, $r(16) = .44$, $p = .07$, although it retained the same approximate strength, indicating that the correlation is not entirely due to the difference between accurate and inaccurate singers. However, in this case, it is inappropriate to remove these cases, since we are interested in precisely these cases of poor singers. There was no significant correlation between melodic singing ability and any measurement of slider pitch-matching ability, nor was there a correlation between any measurements of single-

pitch-matching ability and tonal drift in melodic singing. Melodic singing ability seems to be moderately correlated with vocal pitch matching, but not with slider pitch matching.

To see whether there is a relationship between singing ability and vocal quality, we correlated our measurements of pitch-matching ability (on the slider and voice) and melodic singing ability (mean interval error and tonal drift) with the vocal quality measurements (jitter, shimmer, and pitch stability). We found a significant correlation between the mean interval error during "Happy Birthday" and the internal pitch stability (F_0SD) of the sung target tones, $r(20) = .43$, $p = .05$. Participants with more stable individual tones also produced more accurate melodic intervals, on average. There was also a significant correlation between the tonal drift (measuring the long-term stability of melodic singing) and the target tone jitter, $r(20) = .49$, $p = .02$, and a nonsignificant trend toward a correlation of tonal drift and target tone shimmer, $r(20) = .39$, $p = .07$. In general, participants with a higher voice quality tended to sing "Happy Birthday" more accurately. There were no significant correlations between any vocal quality measurements and slider or vocal pitch-matching abilities.

Discussion

This experiment had two main goals. First, we aimed to discover whether matching a pitch on the slider would be equivalent to matching a pitch with the voice when the timbres were highly similar. Second, we aimed to discover whether single-pitch-matching ability would predict melodic singing ability, in terms of both pitch error and vocal quality.

Slider–voice comparison

Despite the equivalent timbres in the slider and vocal pitch-matching conditions, nonmusicians were still significantly better at matching pitches on the slider than with their voice. Moreover, there was no correlation between participants'
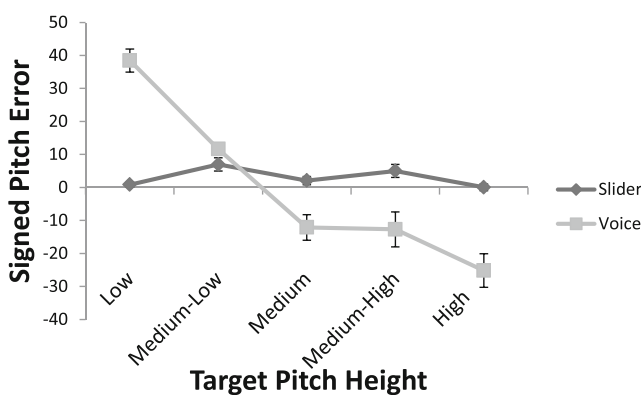

Fig. 2 Signed errors for voice and slider conditions across the five different target heights, with standard error bars
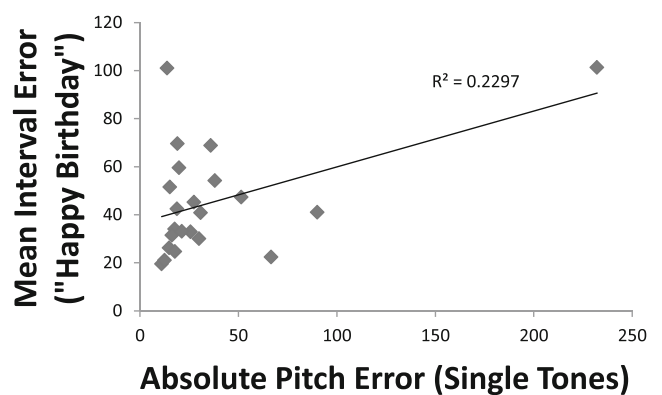

Fig. 3 Scatterplot of mean absolute error in the vocal pitch-matching condition and mean interval error in the melodic singing condition

average errors in each modality. If perceptual problems were a common cause of poor singing, we would expect to see pitch errors on the slider that were similar in size to those in the voice and correlations between these two measurements, due to a fundamental problem of poor pitch perception affecting both subsequent actions. On the contrary, our results confirm that poor singers are able to match the same pitches when using a different motor effector (the finger, in this case). Because there is only one timbre throughout the experiment, for targets and responses, the proximal cause of poor singing here is likely to be poor vocal-motor control (see Hutchins & Peretz, 2012). In addition, the rate of poor singers here (23%) was similar to that found to have a vocal-motor control problem in Hutchins and Peretz (2012) (20%).

Additional evidence for the role of vocal-motor control in pitch-matching tasks can be found in the effects of target pitch height. Although participants were asked to produce target tones in a comfortable range, they showed a significant tendency to sing flat when imitating higher-pitched targets and sharp when imitating lower-pitched targets. There was no similar tendency in responses on the slider, ruling out a perceptually based explanation for this finding. This pattern (which was not present in previous self-matching paradigms; Hutchins & Peretz, 2012) indicates a clear role of vocal-motor control in singing errors; participants had trouble consistently reaching tones that were toward the extremities of their range. The analysis of vocal quality of the self-produced target tones (shown in Table 1) also shows an effect of vocal-motor control on target quality; lower-pitch tones were consistently of lower vocal quality. However, this lower quality did not seem to affect the ability to accurately perceive the pitch of these tones, as evidenced by the consistent, accurate pitch-matching ability across all targets in the slider.

Scaling up to melodies

Another key aim of this study was to find evidence for the utility of pitch-matching tasks in evaluating general singing ability. Although it has been assumed that the ability to match individual pitches is a foundational skill in singing, it had not yet been verified that this specific ability is related to the ability to sing whole melodies in tune. Our results showed a significant correlation between the average amount of error in singing melodies and in vocally matching single tones. Nonmusicians who were better at singing melodies in tune were also better at vocal pitch matching. Furthermore, there was no correlation between pitch-matching ability on the slider and melodic singing ability, indicating that vocal single-pitch matching is measuring a singing-specific ability, rather than a general musical ability. This provides experimental evidence that measuring vocal pitch-matching ability for single tones is a useful tool for gauging overall singing ability. However, it should also be noted that this correlation was not

one to one; there is still a good deal of variance in melodic singing ability unaccounted for in single-pitch matching, some of which may concern tonality. Larrouy-Maestri, Lévêque, Schön, Giovanni, and Morsomme (2013) showed that, together, intervallic error and tonal drift could account for 81% of the variance in experts' judgments of singing ability; our results showing that single-pitch-matching error can predict intervallic error but not tonal drift lend quantitative confirmation to these expert judgments.

Our results also showed a significant correlation between melodic singing ability and vocal quality. People who were better at singing melodies in tune tended to have better vocal quality and sang individual tones more stably. Interestingly, no measurements of vocal quality had any significant correlations with vocal pitch-matching ability. Although this is only correlative evidence, this seems to indicate that the relationship between vocal quality and in-tune melodic singing is not mediated by vocal pitch-matching ability; there is an aspect of melodic singing that is influenced by vocal timbre separately from single-pitch matching. It is also important to note that, because the experimenter selected the tones to use as targets partially on the basis of vocal quality, these represent the upper ranges of quality for each singer.

One possible explanation for the relationship between vocal quality and the ability to sing in tune is that it is likely that those with better vocal quality are also more likely to sing more often; the extra practice may lead to better overall singing ability and better vocal quality. Those who practice singing melodies more often may not practice single-pitch imitation to the same extent, which could be a reason for the lack of correlation between vocal quality and single-pitch matching.

It should also be noted that the aspects of vocal quality measured here (pitch stability, jitter, and shimmer) represent only some possible measurements of vocal quality and timbre. Of these, both pitch stability and jitter, which are different measurements of short-term fluctuation in pitch, are strongly related to the mean pitch. Although these do not directly affect mean pitch in single-tone or melodic contexts, it is less surprising that there would be relationships between these variables (the same does not hold for shimmer, on the other hand, which is a measurement of short-term fluctuation in amplitude). These measurements were chosen because they represent a few standard variables used in assessing vocal quality, but other measurements (e.g., breathiness, spectral centroid, spectral envelope, etc.; Larrouy-Maestri, Magis, & Morsomme, 2014) may reveal a different pattern of effects. In addition, because we measured the vocal quality of only one vowel (/a/), there is the possibility that vocal quality of any given participant may be different across vowels. Exploring the relationship between other aspects of timbre and singing ability would seem to be a fruitful avenue for future exploration.

Conclusion

By using a touch-based measurement of peoples' sensitivity to pitch variations of their own voice, we were able to provide clear evidence that timbral factors are not responsible for the difference between pitch perception and production abilities. Rather, this novel method makes it clearer that vocal-motor problems are a primary cause for many types of singing difficulties. In addition, we have shown that the types of single-tone pitch-matching tasks used here and in many previous studies are a good proxy for melodic singing ability. Finally, differences in timbre and vocal quality can also be associated with the ability to sing melodies in tune.

References

Berkowska, M., & Dalla Bella, S. (2009). Acquired and congenital disorders of sung performance: A review. *Advances in Cognitive Psychology, 5,* 69–83.

Boersma, P. & Weenink, D. (2013). Praat: doing phonetics by computer [Computer program]. Version 5.1.44, retrieved 22 October 2010 from http://www.praat.org/

Bradshaw, E., & McHenry, M. A. (2005). Pitch discrimination and pitch matching abilities of adults who sing inaccurately. *Journal of Voice, 19,* 431–439.

Dalla Bella, S., Giguère, J. F., & Peretz, I. (2007). Singing proficiency in the general population. *Journal of the Acoustical Society of America, 121,* 1182–1189.

de Cheveigné, A., & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America, 111,* 1917–1930.

De Nil, L. F., & Lafaille, S. J. (2002). Jaw and finger movement accuracy under visual and nonvisual feedback conditions. *Perceptual and Motor Skills, 95,* 1129–1140.

Estis, J. M., Coblentz, J. K., & Moore, R. E. (2009). Effects of increasing time delays on pitch-matching accuracy in trained singers and untrained individuals. *Journal of Voice, 23,* 439–445.

Estis, J. M., Dean-Claytor, A., Moore, R. E., & Rowell, T. L. (2011). Pitch-matching accuracy in trained singers and untrained individuals: The impact of musical interference and noise. *Journal of Voice, 25,* 173–180.

Fitts, P. M. (1954). The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology, 47,* 381–391. doi:10.1037/h0055392

Hutchins, S., & Peretz, I. (2012). A frog in your throat or in your ear? Studying the causes of poor singing. *Journal of Experimental Psychology: General, 141,* 76–97.

Hutchins, S., Roquet, C., & Peretz, I. (2012). The vocal generosity effect: How bad can your singing be? *Music Perception, 30,* 147–159.

Larrouy-Maestri, P., Lévêque, Y., Schön, D., Giovanni, A., & Morsomme, D. (2013). The evaluation of singing voice accuracy: A comparison between subjective and objective methods. *Journal of Voice, 27*(2), 259 e1–259 e5.

Larrouy-Maestri, P., Magis, D., & Morsomme, D. (2014). Effects of melody and technique on acoustical and musical features of Western operatic singing voices. *Journal of Voice, 28*(3), 332–230.

Larrouy-Maestri, P., Magis, D., & Morsomme. D. (in press). The evaluation of vocal pitch accuracy: The case of operatic singing voices. *Music Perception*

Larrouy-Maestri, P., & Morsomme, D. (2014). Criteria and tools for objectively analyzing the vocal accuracy of a popular song. *Logopedics, Phoniatrics, Vocology, 39*(1), 11–18.

Moore, R. E., Estis, J., Gordon-Hickey, S., & Watts, C. (2008). Pitch discrimination and pitch matching abilities with vocal and nonvocal stimuli. *Journal of Voice, 22,* 399–407.

Moore, R. E., Keaton, C., & Watts, C. (2007). The role of pitch memory in pitch discrimination and pitch matching. *Journal of Voice, 21,* 560–567.

Pfordresher, P. Q., & Brown, S. (2007). Poor-pitch singing in the absence of "tone deafness". *Music Perception, 25,* 95–115.

Pfordresher, P. Q., & Mantell, J. T. (2009). Singing as a Form of Vocal Imitation: Mechanisms and Deficits. In J. Louhivuori, T. Eerola, S. Saarikallio, T. Himberg, & P.-S. Eerola (Eds.), *Proceedings of the 7th Triennial Conference of European Society for the Cognitive Sciences of Music* (pp. 425–430). Finland: Jyväskylä.

Sataloff, R. T. (2005). *Professional voice: the science and art of clinical care.* San Diego, CA: Plural Publishing.

Titze, I. R. (2000). *Principles of Voice Production. (Second Printing).* Iowa City: National Center for Voice and Speech.

Watts, C., Barnes-Burroughs, K., Adrianopoulos, M., & Carr, M. (2003). Potential factors related to untrained singing talent: A survey of singing pedagogues. *Journal of Voice, 17,* 298–307.

Watts, C., Moore, R., & McCaghren, K. (2005). The relationship between vocal pitch matching skills and pitch discrimination skills in untrained accurate and inaccurate singers. *Journal of Voice, 19,* 534–543.