Journal of Experimental Psychology: General
2012, Vol. 141, No. 1, 76–97

© 2011 American Psychological Association
0096-3445/11/$12.00 DOI: 10.1037/a0025064

A Frog in Your Throat or in Your Ear? Searching for the Causes of Poor Singing

Sean Hutchins and Isabelle Peretz
Université de Montréal

Singing is a cultural universal and an important part of modern society, yet many people fail to sing in tune. Many possible causes have been posited to explain poor singing abilities; foremost among these are poor perceptual ability, poor motor control, and sensorimotor mapping errors. To help discriminate between these causes of poor singing, we conducted 5 experiments testing musicians and nonmusicians in pitch matching and judgment tasks. Experiment 1 introduces a new instrument called a slider, on which participants can match pitches without using their voice. Pitch matching on the slider can be directly compared with vocal pitch matching, and results showed that both musicians and nonmusicians were more accurate using the slider than their voices to match target pitches, arguing against a perceptual explanation of singing deficits. Experiment 2 added a self-matching condition and showed that nonmusicians were better at matching their own voice than a synthesized voice timbre, but were still not as accurate as on the slider. This suggests a timbral translation type of mapping error. Experiments 3 and 4 demonstrated that singers do not improve over multiple sung responses, or with the aid of a visual representation of pitch. Experiment 5 showed that listeners were more accurate at perceiving the pitch of the synthesized tones than actual voice tones. The pattern of results across experiments demonstrates multiple possible causes of poor singing, and attributes most of the problem to poor motor control and timbral–translation errors, rather than a purely perceptual deficit, as other studies have suggested.

Keywords: music cognition, singing, sensorimotor, action, perception

People in modern Western culture are exposed to music quite regularly, either intentionally or inadvertently. Within the set of music that is most popular and listened to, the most important instrument is certainly the human voice. For example, singers are almost always in the forefront of any musical ensemble in which they are involved; shows such as American Idol are incredibly popular around the world; and when people remember songs, it is almost always the vocal part that commands the most attention. So given this exposure to and interest in singing, why is it that a significant percentage of the population fail to sing well?

Of course, singing is a complex and multivariate phenomenon: to say someone sings well is to make a relative judgment about several aspects, such as pitch clarity, voice quality, assertiveness, some that can be easily measured, and others that are more subjective and difficult to measure. What is considered good singing is often a matter of personal preference. However, the most important of these factors in judging singing ability is pitch accuracy. Watts, Barnes-Burroughs, Andrianopoulos, and Carr (2003) asked music educators to identify the factors associated with singing talent and found that the factor cited as being most indicative of a talented singer was intonation—the ability to sing pitches accurately. In addition, among informal singers, the majority of singing errors are pitch errors, rather than timing errors (Dalla Bella, Giguère, & Peretz, 2007), and training in music can significantly aid intonational accuracy (Amir, Amir, & Kishon-Rabin, 2003; Dalla Bella et al., 2007; Murry, 1990; Watts, Murphy, & Barnes-Burroughs, 2003).

Researchers in two studies from 2007 examined the general prevalence of good and poor singing abilities using acoustical measures. Dalla Bella and collaborators (2007) recruited 62 occasional singers to perform "Gens du Pays," a song commonly sung at birthday celebrations in Québec. This song contains 32 tones and 31 intervals. They measured the pitch height of each sung tone and found a wide distribution of singing abilities, as characterized by pitch interval accuracy. Their results indicated that the majority of musically untrained individuals can sing relatively accurately, and there was a range of different abilities across singers. The occasional singers produced few intervals that deviated from the score by more than one semitone, and singing at a slower tempo further aided their pitch accuracy. Pfordresher and Brown (2007)

This article was published Online First August 29, 2011.

Sean Michael Hutchins and Isabelle Peretz, International Laboratory for Brain, Music and Sound Research (BRAMS) and Department of Psychology, Université de Montréal, Montréal, Québec, Canada.

This work was supported by grants from the Canadian Institutes of Health Research, the Natural Sciences and Engineering Research Council of Canada and a Canada Research Chair in Neurocognition of Music to I. P. and a fellowship in Auditory Cognitive Neuroscience from the CREATE program of the Natural Sciences and Engineering Research Council of Canada to S. M. H.

The authors thank Nathanaël Lécaudé and Olivier Piché for technical aid, Catherine Roquet and Marilou Beauchamp-Lalonde for aid in testing participants, and Nathalie Gosselin, Nicolas Robitaille, Sylvie Hébert, Peter Pfordresher, and Isabel Gauthier for their insightful comments on earlier versions of the article.

Correspondence concerning this article should be addressed to Sean Michael Hutchins, Brain, Music and Sound (BRAMS), Pavillon 1430 Boulevard Mont-Royal Montréal, Québec H2V 4P3, Canada. E-mail: sean.michael.hutchins@umontreal.ca

measured intonational accuracy in pitch reproduction tasks among nonmusicians. Participants sang back four-tone sequences of three types: all the same pitch, two pitches (one interval), or four unique pitches comprising a short melody. Eighty-seven percent of the participants sang the sequences in tune (in this case, defined as within one semitone of the targets). These results showed that the poor singers tended to compress the size of intervals in addition to making pitch errors on individual notes. However, the complexity of the stimulus had opposite effects on good and poor singers, such that good singers showed more errors on more complex stimuli, whereas poor singers showed less error for these. Both Pfordresher and Brown (2007) and Dalla Bella et al. (2007) agreed that most people's singing abilities are better than they generally give themselves credit for, and that truly poor singing (as measured by intonational accuracy) is limited to between 10% and 20% of the population. Further studies by Pfordresher and Brown (2009; Pfordresher, Brown, Meier, Belyk, & Liotti, 2010) have confirmed this figure of the prevalence of poor singing.

The nature of the singing task may have some effect on the prevalence of poor singing. For example, certain studies show that singing along with a model may increase singing accuracy (Tremblay-Champoux, Dalla Bella, Phillips-Silver, Lebrun, & Peretz, in press; Wise & Sloboda, 2008), although Pfordresher and Brown (2007) found that the opposite could be true in certain cases, and Hutchins and colleagues (Hutchins, Zarate, Zatorre, & Peretz, 2010) showed no effect of either singing with a model or masking feedback. In addition, there is some evidence that poor singers may be worse at matching single pitches than melodies (Pfordresher & Brown, 2007; Pfordresher & Mantell, 2009), although this topic has yet to be explored in depth.

In the present study, we measured the abilities of musicians and nonmusicians to match single pitches, with the aim of discovering the underlying causes of poor singing abilities, when they occur. We introduce a new method of measuring pitch-matching ability independently of one's ability to control one's own voice. To do so, we used a new instrument called a slider. This instrument is touch sensitive and allows participants to match a pitch instrumentally using a single finger without the requirement for prior instrumental practice. In these studies, we measured pitch perception ability with the slider, and compared it to pitch production ability with the voice, to work out the relationship between the two abilities. This has not been possible in previous studies, which measured perceptual abilities solely with decision tasks. By manipulating the types of stimuli presented to participants and ways in which they can respond, we gained a better understanding of the factors that can cause poor singing.

What Causes Poor Singing?

The wide range of singing abilities is intriguing considering that in speech, the other major use of our phonatory and articulatory systems, almost everybody is able to use their voice adeptly. The National Institute on Deafness and Other Communication Disorders estimates that only about 2.5% of Americans have any type of trouble using their voices (http://www.nidcd.nih.gov/health/statistics/vsl.asp), most of which occurs in children and older adults, and some of which is attributable to identifiable physical or neurological problems. Therefore, it is reasonable to ask why many people fail to sing in tune. Pfordresher and Brown (2007) identified four possible factors that could cause poor singing: a perceptual deficit, a motor deficit, a sensorimotor mismapping, and a memory deficit. We reviewed the evidence for each, as well as the additional factors of motivation and practice .

Perceptual Explanations

One of the most explored hypotheses for poor singing is a deficit in perceptual abilities. A difficulty with ascertaining a target pitch could lead to larger errors when attempting to match it with the voice. According to this explanation, pitch-matching accuracy should be limited by one's pitch discrimination ability. While many studies have correlated general measures of perception ability with vocal pitch matching in neurologically normal participants, only two, Amir et al. (2003) and Nikjeh, Lister, and Frisch (2009), used pitch perception thresholds (just noticeable differences) as a measurement of perception. Although they can have their own flaws or idiosyncrasies (including practice effects and questions about how best to measure them), threshold measurements provide a measurement of perception which uses the same units as production measurements, in contrast to measuring simply the percentage of correct responses in a discrimination task. Amir et al. found a correlation of .67 between perceptual thresholds and vocal pitch-matching accuracy (and this correlation maintained its significance within the nonmusicians, but not the musicians). However, this study had several limitations. Only nine vocalizations per subject were measured, covering three target pitches. Furthermore, the pitch perception thresholds were measured for different pitch ranges than the sung tones, making the comparison between perception and production problematic. This study also indicated that several participants showed good vocal pitch-matching abilities but poor discrimination abilities. Finally, Amir et al. (2003) showed significantly more accurate discrimination than vocal pitch-matching abilities, suggesting that discrimination was not the limiting factor for pitch-matching abilities and that other factors may have been involved.

Nikjeh et al. (2009) improved this design by using complex tones instead of sine wave tones and by equalizing the pitch ranges used for production and discrimination, and showed discrimination abilities on a comparable range to vocal pitch-matching abilities. However, their data showed a relationship between the two abilities only among instrumentalists, but not among trained singers or nonmusicians. In addition, the pitch perception thresholds they measured among all groups were quite high, which they attributed to their procedure and stimulus timbre. Thus, their results left some doubt as to whether their participants' vocal pitch-matching abilities were truly limited by their perceptual discrimination abilities.

Watts, Moore, and McCaghren (2005) found a similar correlation between perceptual abilities and vocal pitch matching among nonmusicians. Instead of measuring just noticeable differences, however, they measured perceptual abilities by assessing pitch discrimination abilities across different timbres for two pitches differing by a perfect fourth (five semitones), and only eight total judgments were made. Vocal pitch-matching abilities were taken from only 24 vocalizations per subject, across eight target tones, which may be too small a sample, given the within-subject variance in vocal pitch matching (Pfordresher et al., 2010). The correlation reported was between the percentage of tone pairs correctly discriminated and the average accuracy of the pitch matching for each subject—two variables that cannot be directly compared. In addition, this study, like Amir et al. (2003) showed that there were several participants with good vocal pitch-matching abilities but poor discrimination abilities. In subsequent studies, Moore and colleagues replicated this correlation, using a similar design (although generally using same timbre comparisons; Estis, Coblentz, & Moore, 2009; Estis, Dean-Claytor, Moore, & Rowell, 2010; Moore, Keaton, & Watts, 2007), although one did fail to find such a relationship (Moore, Estis, Gordon-Hickey, & Watts, 2008). Other studies, however, have found no relationship between pitch discrimination and vocal pitch-matching abilities (Bradshaw & McHenry, 2005; Dalla Bella et al., 2007; Pfordresher & Brown, 2007), among good or poor singers.

The perceptual deficit hypothesis for poor singing posits that pitch discrimination abilities should be the limiting factor for vocal pitch matching. Apart from Amir et al. (2003) and Nikjeh et al. (2009), the studies mentioned previously do not actually measure the limits of perceptual discrimination, but take instead a proportion of correct responses to a same/different judgment task. This latter measurement can be correlated with pitch-matching error but not does not permit the comparison of magnitudes necessary to prove that pitch-matching ability is primarily limited by perceptual discrimination ability.

While these correlations (or lack thereof) can provide useful hints about the relationship between pitch perception and vocal production, it is problematic to make judgments about whether these perceptual abilities *cause* poor pitch singing. For one, most normal subjects tend to have no problems discriminating pitches differing by more than a quarter of a semitone (Hyde & Peretz, 2004), while poor singing is often defined as being mistuned by a half or full semitone. This makes it unlikely that discrimination ability is limiting vocal pitch production ability. In addition, as mentioned earlier, same/different judgment tasks do not provide a measure that can be directly compared with measurements of the accuracy of a sung tone, since they are in different units (percentage correct vs. cents error) and in response to different stimuli (generally, a pair of tones vs. a single tone).

A few studies (e.g., Platt & Racine, 1985) have used knob-controlled set-ups to study musical tuning skills, in which participants adjust a pitch to match a target. This can provide a measurement of tuning accuracy in units of frequency. Demorest (2001) used a similar knob-controlled dial to measure pitch-perception ability among junior high boys and found a relationship between their ability to perform this task and to match a pitch vocally. The junior-high boys, however, had only three trials at the knob with no practice and were allowed to hear the bassoon-timbre target sound continuously, allowing the detection of auditory beating as a possible cue for pitch discrimination. Mean deviations on the dial-tuning task were near 44 cents and varied considerably between the three target pitches. Mean error was counted as zero on the vocal-matching task if within 50 cents of the target, and all other vocal pitch-matching errors were reduced by 50 cents as well. Because of these factors, vocal pitch matching was not precisely measured, and the dial-tuning task likely did not measure a true perceptual threshold, but rather a familiarity with tuning tasks, given the high thresholds obtained over the small number of trials. Demorest and Clements (2007) replicated the general finding with a noncontinuous, computer-based pitch-matching test using an onscreen slider. However, as the response tone was discretized by semitones, this experiment was not designed to look at differences in tuning less than 100 cents, which is well above the perceptual thresholds of virtually all listeners, and instead approximates the task of matching a pitch using a piano. This was done to explore the effect of matching, rather than tuning, but is most likely not sensitive enough to reveal solely perceptual difficulties.

In what is perhaps the most convincing test of the relationship between pitch perception and production abilities to date, Zarate, Delhommeau, Wood, and Zatorre (2010) used micromelody training over six sessions to improve their participants' perceptual abilities. Participants sang and discriminated melodies with very small, nontraditional intervals. Zarate et al. (2010) found no evidence that vocal pitch-matching abilities had improved in the posttraining session, despite their measurable improvement in pitch discrimination.

Despite the lack of consensus about whether perceptual deficits cause poor singing in most poor singers, there is a general consensus that this is the case in congenital amusia. This is a neurodevelopmental disorder that causes severe deficiencies with pitch perception and is often characterized by pitch-poor singing (Ayotte, Peretz, & Hyde, 2002). Congenital amusics have difficulty singing (Ayotte et al., 2002; Dalla Bella, Giguère, & Peretz, 2009; Tremblay-Champoux et al., in press) and matching pitches (Hutchins et al., 2010), and this is thought to be related to their pitch-perception deficits. It is worth noting, though, that some amusics retain fairly good singing and vocal pitch-matching skills, despite their perceptual difficulties. However, this is a relatively rare condition that cannot account for the majority of cases of poor singing.

Motor and Sensorimotor Explanations

Another possible cause of poor singing is a deficit in production abilities. This explanation presumes that poor singers can accurately perceive the pitch they wish to imitate, but lack vocal-motor control necessary to create that pitch. This hypothesis was suggested in early work with school children by Joyner (1969) and Cleall (1970, as cited by Goetze, Cooper, & Brown, 1990). Joyner (1969) posited that "monotone" singing in children was the result of a motor dysfunction, and that improving vocal-motor skills could improve pitch discrimination ability (rather than the other way around). He suggested that this dysfunction may be caused by a music syllabus that was inappropriate for children. Cleall (1970, as cited by Goetze, 1990) showed that vocal range tends to expand with age and posited that the "standard" repertoire was too high for school children, which may cause their poor singing abilities, and that a better understanding of children's actual vocal ranges may aid in their musical development.

Pfordresher and Brown (2007) proposed a similar hypothesis, which they termed a *sensorimotor account*. They posited that poor singing ability, rather than being strictly production- or perception-based, is an imitative problem, with the fault lying in a mismapping between perception and production. They noted that errors in vocal pitch tended to occur in the same direction across changes in target pitches and that feedback did not seem to aid or hinder poor singers. Pfordresher and Brown (2007) ruled out a strictly motor account of their results on the basis of the fact that a subset of the subjects showed no differences in overall vocal range between good and poor singers. A subsequent analysis of these data, presented in Pfordresher and Mantell (2009), also indicated that those categorized as poor singers were not impaired at sustaining level tones compared with good singers. However, motor problems may manifest in other ways than simply a lack of vocal range or vocal stability. Poor ability to accurately control one's vocal apparatus, too, may be a type of motor problem, and poor singers' motor control may simply not be accurate enough to imitate the sequences reliably, despite their physical ability to reach the correct notes (similar to how the physical ability to, for example, throw a bowling ball down the center of the lane does not guarantee that one will bowl a strike each time). In addition, a strictly imitative deficit, without an accompanying motor control deficit, would predict similar variability (but not accuracy) between good and poor singers. In this case, both groups should show a similar amount of deviation around the pitches they sing, whether they are correct or incorrect on average. Prior work has shown that although their errors were more likely to be in a particular direction, nearly all inaccurate singers have a large response variability, and that singers rarely produce a wrong pitch with low deviation (Pfordresher et al., 2010).

Another type of mismapping problem in singing can arise from a confusion between musical timbres. Several experiments have shown that timbre is not perceptually independent of pitch (e.g., Krumhansl & Iverson, 1992; Melara & Marks, 1990a, 1990b, 1990c; Pitt, 1994; Warrier & Zatorre, 2002). Listeners show Garner (1974) interference between classification judgments of timbre and pitch (Krumhansl & Iverson, 1992; Melara & Marks, 1990a, 1990b, 1990c). This effect is strongest for comparisons of isolated tones, and a melodic context seems to aid listeners' ability to perceive pitch independent of timbre (Krumhansl & Iverson, 1992; Semal & Demany, 1991, 1993; Warrier & Zatorre, 2002). Educators have reported that children are more successful in matching pitches vocally when the targets are more like their own voice (reviewed in Goetze, Cooper, & Brown, 1990). This effect has been shown in adults as well (Watts & Hall, 2008), with the most improvement for participants matching recordings of their own voice (Moore et al., 2008). Poor singers may not have a unitary concept of pitch, and might represent the pitch of different instrumental timbres along different, independent dimensions. In this case, their poor singing might arise from their difficulty in "translating" a pitch from the timbre of the stimulus to that of their voice.

Memory and Motivation Explanations

Two other factors may affect pitch-matching accuracy, in some cases. One is a poor memory for pitch. If pitches are perceived accurately, but some error encroached on the memory during its encoding, storage, or retrieval prior to its reproduction, this would lead to a poor vocal pitch match. This explanation was considered and ruled out by Pfordresher and Brown (2007), due to the lack of stimulus complexity effects, although the complexity differences may not have been strong enough to elicit memory-related problems. However, time delays (Estis et al., 2009) and musical or noise interference (Estis et al., 2010) can impair pitch-matching ability, especially in musically untrained individuals.

Finally, poor singing may result from a lack of motivation. Gould (1969) showed that teachers rated motivational factors such as psychological inhibition and lack of motivation as two of the top seven causes of poor singing. Poor singers may simply not be motivated to sing to the best of their abilities, possibly not finding it worth the effort. In addition, motivational factors can also affect people's willingness to practice basic singing skills and can have long-term effects on singing abilities. However, we would like to emphasize that none of these factors attempts to explain the proximal causes of poor pitch singing. While it is likely that many long-term factors can affect singing ability, especially the amount of practice one has, such explanations do not help us understand what is causing the immediate problem of poor pitch matching found in many people. Practice over an extended period of time may increase perceptual ability, or motor coordination, or memory or motivation, but is not by itself an explanation of how any instance of pitch matching has succeeded or failed.

In the current set of studies, we dealt only with the immediate causes of poor singing, that is, the problems that may occur between perception of a musical pitch and its immediate reproduction. Thus, we focused primarily on perceptual, motor, and sensorimotor causes of singing problems, rather than on memory or motivation explanations, which may be more relevant to medium- or long-term singing problems, rather than immediate pitch matching. We tested samples of participants, who would be expected to include a range of singing and musical abilities, especially among nonmusicians, and tested their production ability under different conditions designed to pick out perceptual, motor, and sensorimotor problems when they occurred. Because we could not know in advance whether participants would be good or poor at our tasks, and people often misjudge their own singing ability (Pfordresher & Brown, 2007), we were not able to specifically recruit good and poor singers. Thus, singing proficiency was not included as a grouping factor, but was measured after the fact. Afterwards, we used the varying patterns of results to outline groups of participants whose singing problems varied as a function of the different conditions (or who have no singing problems). In Experiment 1, we examined whether perceptual explanations could account for poor vocal pitch matching, or whether sensorimotor and vocal–motor control explanations are more fitting. In Experiment 2, we aimed to further distinguish between sensorimotor and vocal–motor control explanations and, in Experiment 3, looked at whether the number of attempts could explain performance differences between the slider and vocal pitch matching. Experiment 4 investigated whether lack of visual feedback could explain poor vocal pitch-matching ability. Finally, in Experiment 5, we examined perceptual thresholds for hearing mistuning in vocal and slider sounds, in order to validate some of our analytical choices. The pattern of results across these experiments, and especially within the three conditions presented in Experiment 2, elucidate the specific perceptual, sensorimotor, and vocal–motor problems.

Experiment 1

In the current set of studies, we aimed to more fully understand why some singers have a difficult time matching pitches vocally. In Experiment 1, we examined the link between perception and production in an innovative way and endeavored to separate out perceptual explanations from sensorimotor and vocal–motor control explanations. We used the slider to provide a nonvocal pitch-matching measurement that could be directly compared with vocal pitch matching. The slider was designed to be as similar to the voice as it could be, while using a different motor control mechanism. The timbre of the slider was modeled on the voice, including formants (resembling the syllable /a/) and a moderate amount of vibrato. In addition, the slider is not quantized but can play a continuously changing pitch across one linear dimension, similar to how the human voice can change continuously in pitch across the dimension of vocal fold tension (see Sundberg, 1987, for a detailed discussion of how pitch variations are produced by the vocal mechanism), and both can only sound one pitch at a time. Just as the tension of the vocal folds can be changed without engaging the voice, one can change finger position along the slider without pressing down to engage the sound, to allow responses to be prepared before actually engaging them. Both the slider and voice require some fine motor control, from parts of the body (finger and vocal folds, respectively) that have the requisite amount of fine motor control. The arm–hand–finger system, in particular can make adjustments as fine as .1 mm (De Nil & Lafaille, 2002; Fitts, 1954), thus ensuring that precision of motor control is not a limiting factor in using the slider. Because of this, the ability to accurately perceive the target note should be the main limitation of accuracy of pitch matching using the slider. Furthermore, the similarity of the slider and vocal matching allows us to use the same experimental paradigm across both tasks. The stimuli, methodology, and analyses are the same across both voice and slider conditions; the only change is in the mode of the response. Therefore, if response accuracy differs across the two response modes, it is likely to be related to the sensorimotor or motor attributes of the response mechanism.

Because of these congruencies, performance on slider pitch matching could help to explain the root causes of poor vocal pitch matching. If participants tended to match pitches better with the slider than with the voice, then this would indicate that their perception is not the limiting factor for their vocal pitch-matching abilities and that any poor singing is due to a sensorimotor deficit or a lack of vocal–motor coordination. On the other hand, if performance on the slider was the same as or worse than vocal pitch matching, then this would provide evidence that the vocal apparatus could be just as accurate as the arm–hand–finger system, which is already known to be quite accurate (De Nil & Lafaille, 2002; Fitts, 1954), and any poor singing that is accompanied by poor slider performance likely would be caused by poor perceptual abilities.

We hypothesized that we would find a range of abilities for vocal pitch matching, among both musicians and nonmusicians, but that each subject would perform well on the slider pitch-matching task. Typical just-noticeable differences for pitch are estimated to be near 5 cents in the fundamental frequency range of most singing (Zwicker & Fastl, 1999, although this may be higher for less trained listeners); we hypothesized that slider performance would reflect these difference limens but not the voice performance. These results would support the production-deficit hypothesis. In addition, we also expected to find better performance in both tasks by musicians than nonmusicians, with a more pronounced difference in the vocal task than the slider task.

Method

Participants. Participants were 25 nonmusicians (15 women, 10 men) and 13 musicians (six women, seven men) recruited from the Université de Montréal population. Nonmusicians all had 1 year or less of formal training ($M = 0.2$ years), defined as private, individual lessons on an instrument or the voice, and ranged in age from 18 to 30 (mean age = 23.1 years). Nonmusician participants reported a mean of 0.4 years of group singing experience and no formal singing training. Musicians all had 7 years or more of formal training on their primary instrument ($M = 11.6$ years) and ranged in age from 19 to 28 (mean age = 22.9 years). Musician participants reported a mean of 3.5 years of group singing experience and a mean of 0.6 years of formal singing training. No subjects reported any diagnosed hearing deficits or neurological disorders.

Stimuli and equipment. Target tones were complex waves, made to approximate the timbre of a human voice on the syllable /a/, and were presented to participants in Max/MSP (Cycling '74, San Francisco, CA) through DT 990 Pro headphones (Beyerdynamic, Heilbronn, Germany). FM vibrato with a frequency of 5 Hz and an amplitude of ± 7 cents was added to these tones to better approximate the timbre of the human voice. A Fourier analysis of a sample target tone and a subject's vocal response is shown in Figure 1. Five target tones were used: B3 (246.94 Hz), C#4 (277.18 Hz), D#4 (311.13 Hz), F4 (349.23 Hz), and G4 (392.00 Hz) for women and one octave below for men. These tones are all one whole step apart and do not evoke any single key or tonality. Participant responses were made either vocally, recorded with a TLM 103 microphone (Georg Neumann, Berlin, Germany), or with a simple instrument called a slider. The slider was made from two sensors, a pressure sensor and a 50-cm position sensor (Infusion Systems, Montreal, Québec, Canada). The position sensor was placed over the pressure sensor and mounted on a hard surface between two 1/8-in. rails. When pressed by a finger with sufficient force, the slider sent a 10-bit midi signal indicating the position of the pressure applied. This signal was read by Max/MSP and converted into the same type of complex wave used for the target tones. The fundamental frequency (F_0) of this wave was a function of the slider's position, such that for women, $F_0 = 220 * 2$ ˆ (Position/1023), where position ranged from 0 to 1023. For men, the frequencies of the slider were halved (down one octave) to better approximate the male vocal range. This produced a slider with a range of one octave, between A3 (220 Hz) and A4 (440 Hz) for women and between A2 (110Hz) and A3 (220 Hz) for men, with an equal-temperament output, where all semitones are equally distant. The slider was quantized with steps of 1.17 cents, approximately 85 steps within each semitone (making it not technically continuous, but an approximation of a continuous output, similar to how a television or movie is an approximation of a continuously changing picture). One semitone spanned a physical distance of 4.17 cm on the slider, and each step was approximately .5 mm apart. Although this is a small distance, people were able to make movements with an accuracy of up to .1 mm using their fingers (De Nil & Lafaille, 2002; Fitts, 1954). Because of this, the primary limitation on slider performance should be perceptual, rather than a motoric, limitation.

Procedure. The experiment was divided into two sections, vocal pitch matching and instrument-based pitch matching. We performed the instrument-based pitch-matching section first, so that participants would not be tempted to use strategies learned in the vocal pitch-matching section in the other section (because the slider is a new instrument to participants, we considered it unlikely that they would apply slider-based strategies to the vocal-matching

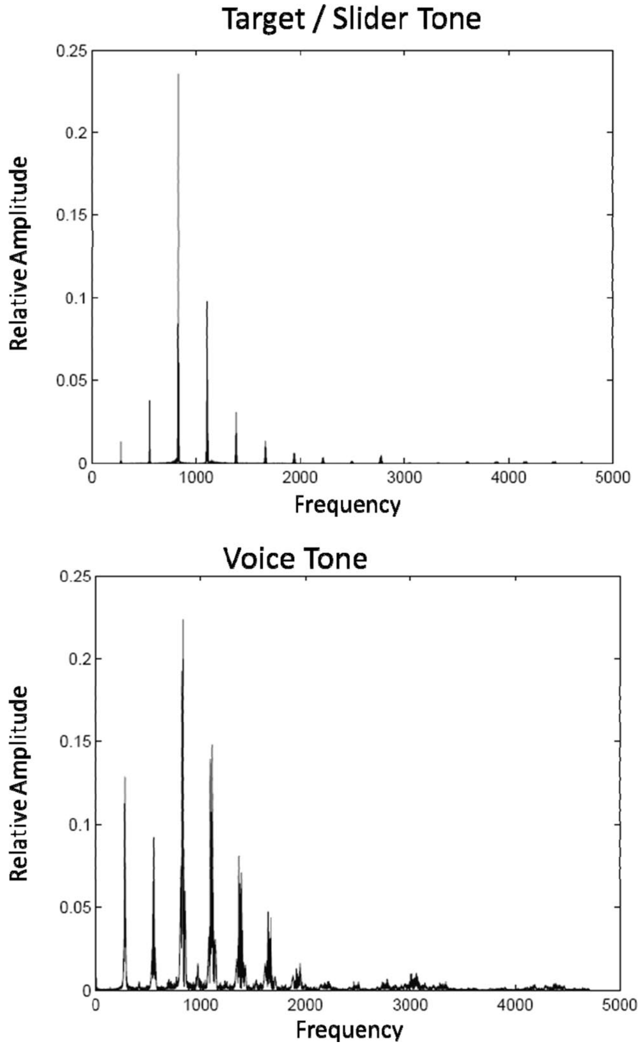

Figure 1. Fourier analyses of a sample target tone and a voice tone, both at 277 Hz (C#4).

task). Both sections used the same trial design and differed only in the medium of the response. Each trial was initiated by the participant pressing the space bar, followed by a continuous presentation of the target tone. Participants were instructed to match the target tone as closely as possible using either the slider or the voice, depending on the section. In order to prevent pitch matching through hearing the beating and acoustical dissonance between the target tones and the tones produced with either the voice or slider, the target was turned off whenever the slider was being used or a vocal response was being made. The target was turned back on when the slider no longer sensed any input or the vocal response was discontinued. Participants were told that they could take as long as they liked to match the target and that their accuracy would be judged only by their final response.

Each section was composed of 100 trials, with 20 instances of each of the five target tones, presented in pseudorandom order, such that the same tone was never repeated immediately. Due to time limitations, not all subjects could complete the full 100 trials in each section; however, each subject performed trials in both slider and voice conditions. The outputs of the slider were saved as .txt files reporting the target tones and the frequency of the slider at each millisecond interval, and participants' voices were recorded as .aif files, allowing all data from each trial to be recorded in full. Prior to the slider pitch-matching subsection, the slider was introduced to the participants, and they were given time to use it freely to produce sounds and become acquainted with it. Vocal warm-ups, including self-directed vocal slides, single-tone matching, and simple melody imitation, were conducted before the vocal pitch-matching subsection. Both subsections were preceded by five practice trials using different targets pitches than in the main experiment.

Results

Data analysis. Prior to further analysis, the vocal recordings were analyzed with a Matlab (The MathWorks, Natick, MA) implementation of YIN (de Cheveigné & Kawahara, 2002), which provided information about frequency, amplitude, and aperiodicity at a rate of 1378 Hz. Responses in both vocal and slider pitch-matching conditions tended to consist of multiple instances of discrete tones, without much change in pitch over a single tone. Both types of responses were analyzed for the same properties. We measured the pitch of the initial tone produced, the pitch of the final tone produced, the total number of discrete responses in each trial, and the total time spent in each trial. All pitch information was converted to cents (1 semitone = 100 cents) in order to make meaningful comparisons between different tones. In order to avoid sharp and flat errors canceling each other out, we used the absolute values in the analysis of the errors. Final responses in each task were considered accurate if the pitch was within 50 cents (1/2 semitone) of the target, a criterion that we subsequently validated in Experiment 5.

Pitch matching. Both musicians and nonmusicians showed a number of differences between the slider and voice conditions. Both groups performed better with the slider than with their voice, and musicians were better overall in both conditions. Because performance with both the slider and voice can be characterized in many ways, we measured a number of different variables. These included the average absolute value of the pitch error during the initial tone produced (to measure initial error), the average absolute value of the pitch error during the final tone produced (to measure final error), the proportion of accurate final responses, the average trial duration, and the average number of discrete responses for each subject in both vocal and slider conditions. Figure 2 shows the average values for each of these measurements.

We performed five separate 2 × (2 × 5) mixed-design analyses of variances (ANOVAs) over the factors of music experience (musician or nonmusicians), response modality (voice or slider), and target pitch, using each of these dependent variables (all values reported are Greenhouse–Geisser corrected, which adjusts the degrees of freedom). All five ANOVAs revealed significant main effects of response modality and music experience, and all the ANOVAs except the one using initial response error also showed a music experience by response modality interaction. None of the measurements showed any effect of target pitch or any higher level interactions with target pitch.

Initial response error. Participants' initial responses were closer to the target in the voice condition than in the slider

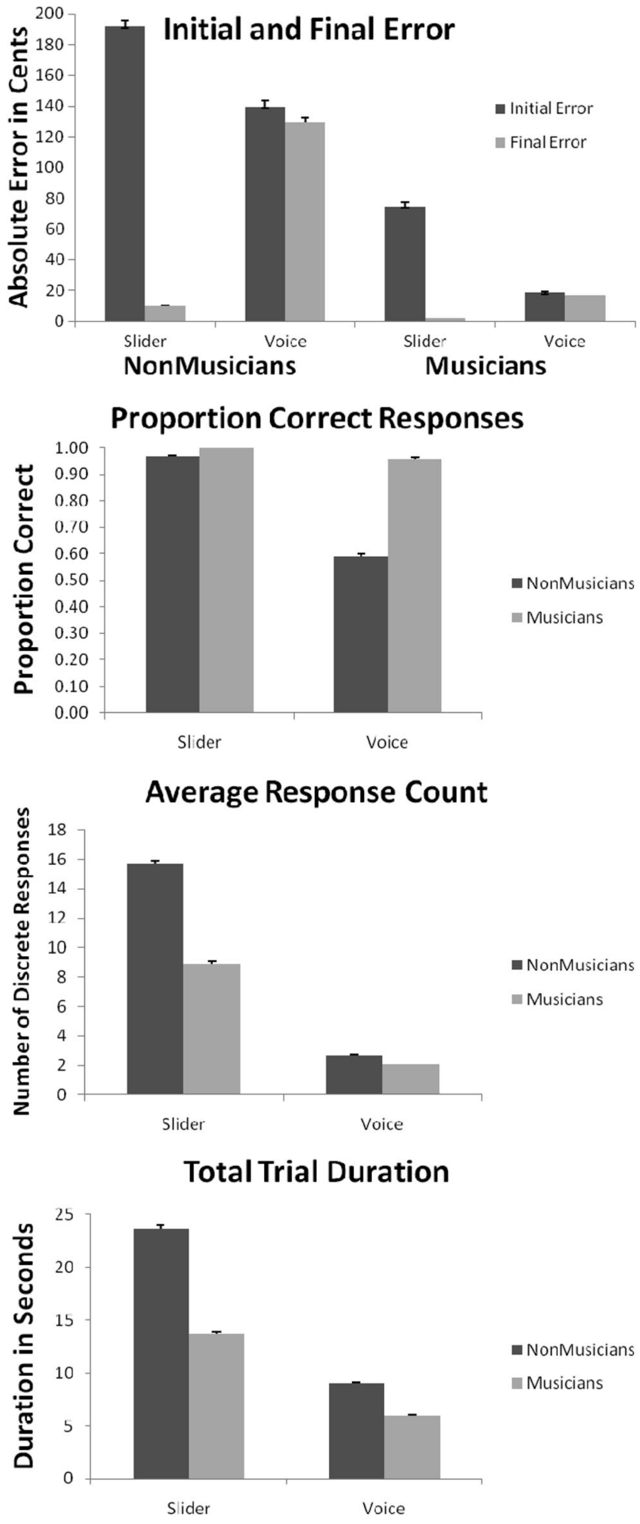

Figure 2. The average absolute value of the pitch error during the initial and final tone produced, the proportion of accurate final responses, the average number of discrete responses, and the total trial duration, shown for musicians and nonmusicians in both vocal and slider conditions from Experiment 1, with standard error bars.

condition, $F(1, 36) = 6.94$, $p = .012$, $\eta^2 = .10$. Musicians' initial responses were closer to the target than those of nonmusicians, $F(1, 36) = 18.04$, $p < .001$, $\eta^2 = .33$.

Final response error. Participants' final responses were closer to the target (as measured in cents) in the slider condition than in the voice condition, $F(1, 36) = 11.76$, $p = .002$, $\eta^2 = .18$. Musicians' final responses were closer to the target than those of nonmusicians, $F(1, 36) = 8.34$, $p = .007$, $\eta^2 = .19$; the difference between musicians and nonmusicians was significantly larger in the voice condition than the slider condition, $F(1, 36) = 7.21$, $p = .011$, $\eta^2 = .11$.

Proportion of correct responses. Participants made more correct responses (i.e., final responses within 50 cents of the target pitch) in the slider condition than in the voice condition, $F(1, 36) = 17.18$, $p < .001$, $\eta^2 = .23$. Musicians made more correct responses than nonmusicians, $F(1, 36) = 14.56$, $p = .001$, $\eta^2 = .29$, and the difference between musicians and nonmusicians was significantly larger in the voice condition than the slider condition, $F(1, 36) = 11.20$, $p = .002$, $\eta^2 = .15$.

Response count. Participants made more discrete responses per trial in the slider condition than in the voice condition, $F(1, 36) = 109.19$, $p < .001$, $\eta^2 = .53$. Nonmusicians made more responses per trial than musicians, $F(1, 36) = 11.99$, $p = .001$, $\eta^2 = .25$, and the difference between musicians and nonmusicians was greater in the slider condition than in the voice condition, $F(1, 36) = 10.94$, $p = .002$, $\eta^2 = .05$.

Total trial duration. Participants spent more time per trial in the slider condition than in the voice condition, $F(1, 36) = 76.76$, $p < .001$, $\eta^2 = .15$. Nonmusicians spent more time per trial than musicians, $F(1, 36) = 10.81$, $p = .002$, $\eta^2 = .22$, and the difference between musicians and nonmusicians was greater in the slider condition than in the voice condition, $F(1, 36) = 7.533$, $p = .009$, $\eta^2 = .03$.

As a follow-up to this, we computed the absolute value of the difference between the initial pitch and final pitch produced, as a measurement of how much the response changed in each trial upon receiving aural feedback. Target tone was omitted from this analysis, as it had no effect on initial or final pitch matching, making this a 2 (music experience) × 2 (response modality) ANOVA. There was a significant main effect of response modality, with considerably more adjustment made in the slider condition than in the voice condition, $F(1, 36) = 86.82$, $p < .001$, $\eta^2 = .63$. Nonmusicians made more adjustment than musicians overall, $F(1, 36) = 22.36$, $p < .001$, $\eta^2 = .38$, and the difference between nonmusicians and musicians was larger in the slider condition than in the voice, $F(1, 36) = 14.99$, $p < .001$, $\eta^2 = .11$. Other follow-up analyses showed no differences in accuracy between genders. There was a strong correlation between mean final response error and the standard deviation of the final response error in the vocal task, $r(36) = .888$, $p < .001$, and between the same measurements in the slider task, $r(36) = .947$, $p < .001$.

To further ascertain how participants were adjusting their responses during a trial, we also analyzed the total number of pitch changes during a trial that were in the incorrect direction, (i.e., instances where their response moved further from the target). Only pitch changes larger than 25 cents (~1 cm on the slider) were considered for this analysis, to ensure that small tuning variations, which may reflect motor variability, were not counted. This measure was used rather than the proportion of responses moving in the correct direction to avoid a dividing by zero error in cases where participants made only one response or their response never changed in frequency during a trial, which was frequent in the voice condition, as can be seen in Figure 2. A 2 × 2 mixed-design ANOVA including the factors of music experience and response modality showed significantly more pitch changes in the incorrect direction in the slider condition than in the voice condition, $F(1, 36) = 15.25$, $p < .001$, $\eta^2 = .24$. Nonmusicians made more pitch changes in the incorrect direction than musicians, $F(1, 36) = 14.13$, $p = .001$, $\eta^2 = .28$, and the difference between nonmusicians and musicians was greater in the slider condition ($M = 1.61$ for nonmusicians, .07 for musicians) than the voice condition ($M = .27$ for nonmusicians, .02 for musicians), $F(1, 36) = 12.94$, $p = .001$, $\eta^2 = .20$.

We also examined the correlation between the final pitch produced and the target pitch for musicians and nonmusicians for both the slider and voice modalities. Figure 3 shows the average pitch matching performance in both the slider and voice modalities, as well as the correlation between targets and both slider and voice performance for nonmusicians and musicians, respectively. In nonmusicians, the slider performance showed no tendency toward either sharp or flat errors, $t(24) = 1.60$, *ns*, $d = 0.33$. However, the voice performance showed errors skewed significantly in the negative (flat) direction, $t(24) = 2.35$, $p = .03$, $d = 0.47$. Musicians showed the same trend, with no dominant direction for error in slider performance, $t(12) = 1.39$, *ns*, $d = 0.27$, and a trend towards flat errors in the voice condition, $t(12) = 1.87$, $p = .09$, $d = 0.52$.

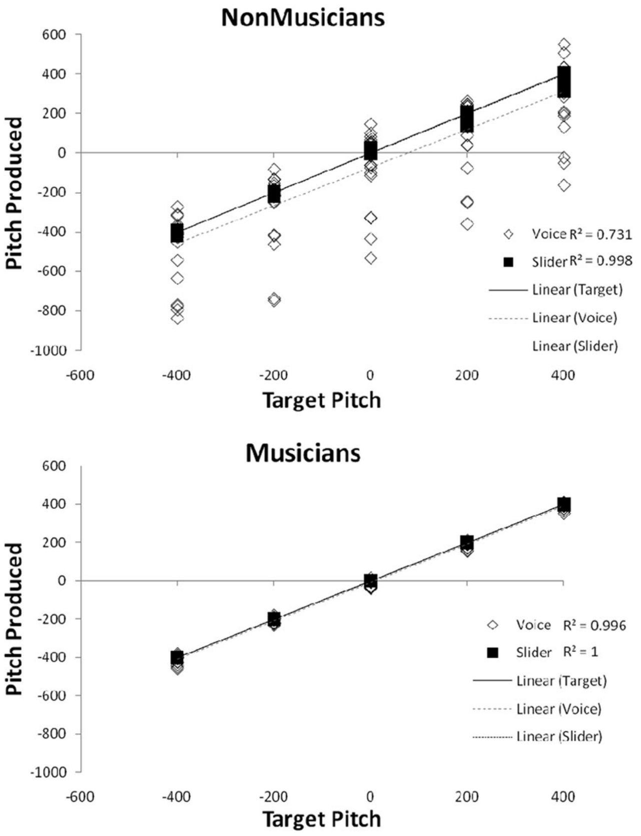

Figure 3. The average pitch matching performance for each subject in both the slider and voice modalities, as well as the correlation across all subjects between targets and both slider and voice performance, shown for each nonmusicians and musicians from Experiment 1. Perfect performance is shown as the solid line; linear regressions for musicians and nonmusicians are shown by dashed lines.

Finally, we examined the correlations between the voice and slider performance, using the proportion of accurate final responses and the absolute values of the pitch errors during the final tone produced as measurements. Accuracy measurements showed no correlation between voice and slider performance, $r(36) = .21$, *ns*; however, the pitch error measurements showed a significant correlation between these two modalities, $r(36) = .50$, $p = .001$ (see Figure 4). Participants who were less precise in using the slider were in general less precise in matching pitches with their voice, although the absolute error was much greater in the voice modality. However, when broken down into musician and nonmusician groups, this correlation only held among nonmusicians, $r(23) = .40$, $p = .05$, and was absent among musicians, $r(11) = -.03$, *ns.* As a follow-up, we compared mean slider error among good singers and poor singers (those whose vocal errors were less than or more than 50 cents, respectively). As would be expected from the significant correlation presented earlier, good singers were more precise on the slider than were poor singers, $t(18) = 2.64$, $p = .02$, $d = 0.94$. However, when musicians (who were all good singers) were removed from the sample so that only nonmusicians were tested, there was no effect of singing ability on slider precision, $t(23) = 1.64$, *ns*, $d = 0.65$.

Discussion

The results of this study show that both musicians and nonmusicians can accurately match pitch using the slider. Overall, participants showed much greater accuracy in matching pitches with the slider than with their voices. Musicians showed better performance overall, but the majority of the difference was in the voice condition, rather than the slider condition. Musicians were nearly perfect in matching pitches with the slider, making no errors whatsoever, and matching the target pitch to within 2 cents on average. This accuracy was consistent across musicians with different main instruments, including not only string and wind instruments, which require similar types of efforts for purposes of tuning, but also piano and drums, which are rarely tuned by those who play them. This indicates that musicians' superior performance on this task is probably not simply due to their greater familiarity with instrumental tuning in general.

Nonmusicians were also highly accurate on the slider task. As a group, their accuracy was near ceiling, making only 3% errors. The average tuning mismatch of all nonmusicians was only 10 cents, or one tenth of a semitone, which is smaller than many people's just noticeable differences, and 10 of the 25 nonmusicians (40%) were as accurate in this task as the musicians, with average errors less than 4 cents. In fact, only three of the 25 nonmusicians (12%) showed average mismatches above 20 cents, and the lowest performing of these still had an average error less than the 50 cent limit and successfully matched the pitch to within that limit on 69% of the trials. Where nonmusicians show markedly different performance from musicians is in how they get to their accurate pitch matches. Nonmusicians spent an average of 23.63 s on each trial, making 15.67 different responses on average each time. This

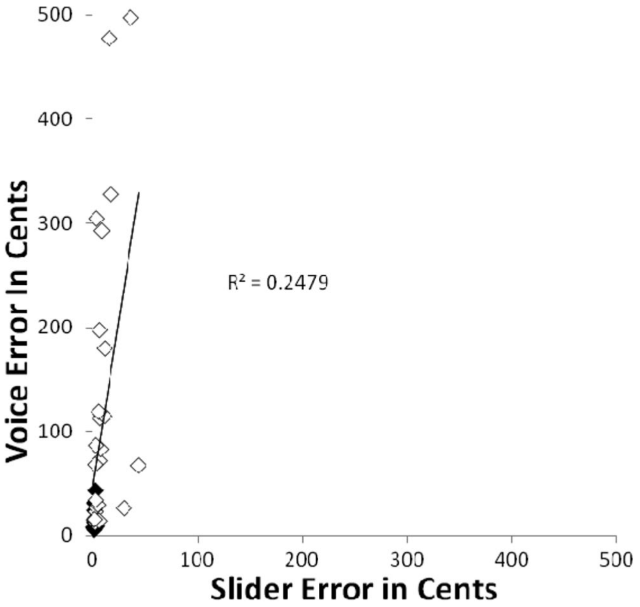

Figure 4. The average absolute slider error plotted against the absolute voice error (both in cents) for musicians and nonmusicians from Experiment 1. Musicians are represented by the black diamonds and nonmusicians by the white diamonds.

is considerably more time spent per trial than the musicians, among whom only one spent even above 20 s per trial, and only three of whom made even 10 slider responses per trial. Partially, this is due to the nonmusicians' greater initial error on the slider. However, we also found that nonmusicians made considerably more errors in how they adjusted their slider responses to hone in on the target, making 1.61 tuning adjustments in the wrong direction on average. Musicians, on the other hand, made very few tuning adjustments that did not bring them in the direction of the target pitch. This reflects their greater fluency with the linear dimensionality of pitch; musicians can more easily discriminate whether their slider response was too high or too low and adjust it accordingly.

In comparison, there was a considerable amount of variability in the voice condition. Among musicians, accuracy was pretty good, but not perfect as with the slider. Musicians matched the pitch accurately with their voice 96% of the time, with three of 13 making more than one error. Musicians' mean error was 17 cents. It is notable that the one musician whose main instrument was the voice was not the most accurate, placing as only the sixth most accurate of the 13.

Nonmusicians spanned the gamut of singing abilities, including some who were just as good as the musicians and some who could not match pitches with their voice. Overall, they were less accurate than the musicians, accurately matching only 59% of the target pitches, with a mean error of 129 cents, well over a semitone (although this number is unduly affected by the poorer singing; the median error for nonmusicians is only 72 cents). When they occurred, errors tended to be flat, rather than sharp, possibly reflecting the greater effort required to sing higher pitches (although this may be a general effort issue rather than a range issue; errors were no more likely on higher target pitches). These tones were not out of range, though, as all participants demonstrated that they were able to sing high enough to reach the target pitch in vocal warm-ups. Ten of the 25 nonmusicians (40%) showed vocal pitch-matching abilities similar to those shown by the musicians, whereas another 10 (40%) showed average errors over 1 semitone. These latter 10 subjects all made errors on at least 65% of trials, with three of those participants successfully matching less than 5% of targets. However, despite the general performance differences between musicians and nonmusicians at this task, the way in which they approached the task was more similar than in the slider condition. Although musicians spent longer in the voice task than nonmusicians, this difference was less than in the slider condition. Furthermore, musicians made only 0.60 more responses per trial on average than nonmusicians.

There was a correlation between pitch-matching abilities in the slider and voice conditions, similar to what has been shown in other studies (Amir et al., 2003; Watts et al., 2005). In the present study, the correlation was not as strong, and like Amir et al. (2003), did not retain its significance among musicians. However, the present study allowed us to directly compare instrumental-based pitch matching and vocal pitch matching. As Figure 4 makes clear, perceptual ability, as measured by the slider task, is not a major limiting factor in vocal pitch matching. In fact, only one subject showed more accurate pitch matching with the voice than with the slider as measured in final pitch error, and this was only by 3 cents, which was within 1 standard error. Many subjects showed poor singing abilities but quite good perceptual abilities, a pattern that has been documented in a few other studies as well (e.g., Dalla Bella et al., 2007; Pfordresher & Brown, 2007). Thus, despite the moderate correlation found here, poor perceptual ability is likely not the cause of these subjects' poor singing abilities. In Experiment 2, we looked further at nonmusicians to pick out potential causes of pitch errors in the voice condition and separate out whether these pitch-matching errors are caused by a vocal–motor control deficit or a sensorimotor deficit.

Experiment 2

One point of difference between the slider and voice conditions in Experiment 1 concerns the timbre of the target tones. In the slider condition, participants matched tones of exactly the same timbre, whereas in the voice condition, although the timbres were similar, the timbre of any participant's sung responses was different from the target tone. Experiment 2 added a third response condition in which participants first recorded their own voice at different pitch levels and then attempted to match their own voices as targets. This ensured that the timbres of the target and the response were as similar as possible. If participants can match their own vocal targets but not the synthesized vocal tones, this is evidence that they have no problems with their vocal–motor control but do have a sensorimotor problem, translating between the synthesized voice timbre and their own voice's timbre. However, if participants fail to match either type of target, this is evidence that the problem lies primarily with poor vocal–motor control. To compare our slider results with more standard procedures of pitch perception, we also measured participants' pitch discrimination thresholds. Only nonmusicians, the group that showed difficulty with the voice task, were tested in this experiment.

Method

Participants. Forty nonmusicians were recruited from the Université de Montréal population for a study involving two sessions. Nine participants failed to show for the second session and were cut from the analyses, yielding 31 participants (18 women, 13 men). Three of the participants had previously participated in Experiment 1 and therefore only participated in the second session of the current experiment. Participants all had less than 2 years of formal training (*M* = .40 years) and ranged in age from 19 to 30 (mean age = 22.68 years). Participants reported a mean of 0.60 years of group singing experience and no formal singing training. No subjects reported any diagnosed hearing deficits or neurological disorders.

Stimuli and equipment. Stimuli and equipment were the same as in Experiment 1, and examples of the participant's own voice were recorded during the experiment and used as target stimuli as well. Auditory pitch thresholds were tested with maximum likelihood procedure (MLP; Grassi & Soranzo, 2009), implemented through Matlab (The MathWorks).

Procedure. Experiment 2 comprised two separate testing sessions. The first session was identical to Experiment 1, including both slider and vocal pitch matching sections. The second session was divided into three subsections. In the first subsection, participants recorded three different examples of five different sung tones, all on the syllable /ba/, for a total of 15 total recordings. Participants were instructed to sing on a low tone, a medium-low tone, a medium tone, a medium-high tone, and a high tone. All tones were self-selected to ensure that the participants chose tones that were comfortably in their own voice range. Each tone was recorded on a separate track and had a duration of about 2 s. Participants were instructed not to choose tones that were at the extreme limits of their range and to ensure that the tones they chose were noticeably different from one another. After this, the experimenter chose the best example from each of the five categories, using the criteria of pitch stability, differentiability from other pitches, and vocal quality. These were then normalized and trimmed to remove silence and used as the five target tones in the second subsection. The second subsection was identical to the vocal pitch-matching section of the first session, with the exception that the target tones were played only once initially. Participants were instructed to wait until the self-recorded target tones had finished playing before singing, and they could choose to hear the target tone again by pressing the *Enter* key. Finally, in the third subsection, participants' pitch discrimination thresholds were estimated using the MLP program (Grassi & Soranzo, 2009). This experiment was performed in two parts, the first estimating pitch discrimination threshold and the second estimating the ability to detect whether a pitch is higher or lower. In each part, participants heard two tones with the same timbre as the target tones and slider from Experiment 1. Each tone lasted 1 s, and the two tones were separated by 0.25 s. Both parts comprised three blocks of 24 trials each. In the first session, participants answered whether the two tones were the same or different. In 20% of the trials, tones with the same pitch were used, as catch trials. In the second section, participants answered whether the second tone was higher or lower in pitch, and all trials were different pitches. Thresholds were determined at the end of the session by the MLP (Grassi & Soranzo, 2009; Green, 1990, 1993). The entire session lasted a little under 1 hr, and most participants finished all of the trials in this session.

Results

Data were analyzed in the same way as in Experiment 1, and we computed the pitch of target tones in the self-matching section using the same method used to compute the pitch of the final produced tone in the vocal-matching section. Figure 5 shows the pitch of the target tones generated by participants in the self-matching section. Overall, participants chose target pitches in the self-matching section ($M = -275$ cents) lower than the experimenter-defined target pitches of the slider or voice sections ($M = 0$ cents), $t(253) = 6.27$, $p < .001$, $d = 0.71$. The distribution of target pitches was significantly skewed to the right (skewness = .631, *SE* = .195), indicating a longer tail on the right (sharp or high) side of the distribution. The target tones chosen by men and women did not differ on average, after the octave adjustment, nor did their ranges or skewness. Participants' mean target skewness was .292, significantly greater than 0, $t(30) = 3.35$, $p = .002$, $d = 0.60$, indicating that this effect was not solely an effect of combining data across subjects. In the pitch discrimination threshold tasks, participants showed a mean threshold of 18 cents for same/different judgments, and a mean threshold of 25 cents for higher/lower judgments. A paired-sampled *t* test showed that these two thresholds were not significantly different from each other, $t(30) = 1.41$, *ns*, $d = 0.32$.

In the pitch-matching tasks, participants showed differences in all three conditions in both their accuracy and how they performed the task. Overall, participants performed best in the slider condition and worst in the original voice condition, with the self-matching condition falling between the two. Figure 6 shows the average values of participants' initial pitch error, final pitch error, proportion of correct responses, number of distinct responses, and total trial duration. For each of these measurements, a repeated-measures ANOVA was performed with the factor of response modality (slider, voice, self-matching). All five ANOVAs revealed significant main effects of response modality and music experience, and paired-samples *t* tests were perform to compare each pair of response modalities, with a Bonferroni correction applied (adjusted α for 15 comparisons = .003). All dependent variables showed significant differences from each other in each condition, with the sole exception that the initial pitch error in the slider condition was not different from the initial pitch error in the voice condition.

Initial response error. Participants' initial response errors were different across response modalities, $F(1, 38) = 16.41$, $p < .001$, $\eta^2 = .35$. Participants were initially more accurate in the self-matching condition than in the voice condition, $t(30) = 3.74$, $p = .001$, $d = 0.67$, or the slider condition, $t(30) = 11.15$, $p < .001$, $d = 2.00$, but the slider and voice conditions did not differ from each other, $t(30) = 0.784$, *ns*, $d = 0.14$.

Final response error. Participants' final responses errors were different across response modalities, $F(1, 31) = 18.79$, $p < .001$, $\eta^2 = .39$. Participants were more accurate in the slider condition than in the voice condition, $t(30) = 4.66$, $p < .001$, $d = 0.84$, or the self-matching condition, $t(30) = 5.12$, $p < .001$, $d = 0.92$, and were more accurate in the self-matching condition than the voice condition, $t(30) = 3.94$, $p < .001$, $d = 0.71$.

Proportion of correct responses. The proportion of correct responses was different across response modalities, $F(1, 39) = 29.87$, $p < .001$, $\eta^2 = .50$. Participants made more correct re-

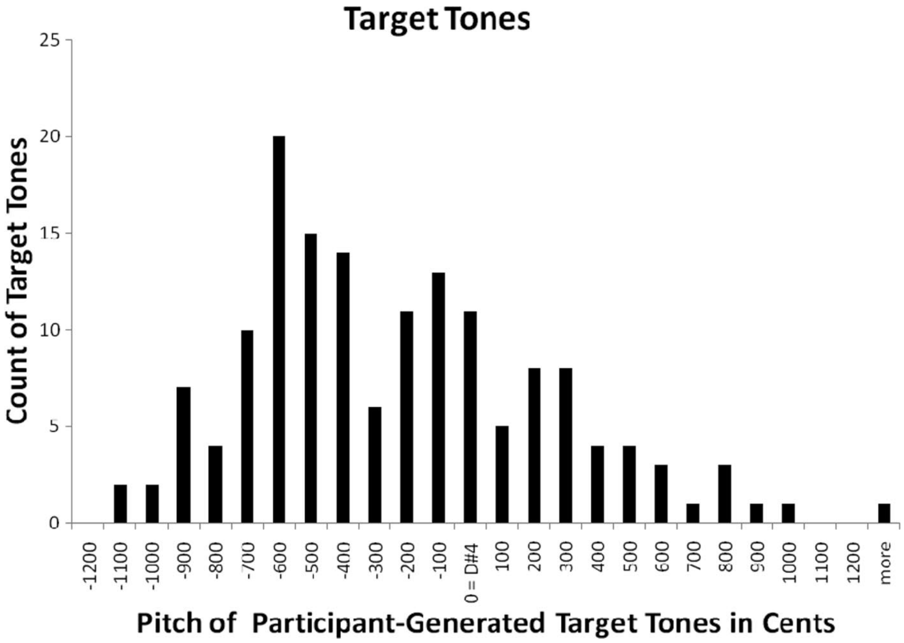

Figure 5. Histogram of the target tones chosen in the self-matching section from Experiment 2, sorted by frequency, in bins of 100 cents. D#4 (311 Hz) is represented as 0, as it is the middle of the slider. Examples from men are transposed up an octave.

sponses in the slider condition than in the voice condition, $t(30) = 5.87$, $p < .001$, $d = 1.05$, or the self-matching condition, $t(30) = 3.25$, $p = .003$, $d = 0.58$, and made more correct responses in the self-matching condition than the voice condition, $t(30) = 5.35$, $p < .001$, $d = 0.96$.

Response count. The number of discrete responses in each trial was different across response modalities, $F(1, 36) = 82.59$, $p < .001$, $\eta^2 = .72$. Participants made more responses in the slider condition than in the voice condition, $t(30) = 9.59$, $p < .001$, $d = 1.56$, or the self-matching condition, $t(30) = 9.33$, $p < .001$, $d = 1.63$, and made more responses per trial in the voice condition than the self-matching condition, $t(30) = 3.51$, $p = .001$, $d = 0.94$.

Total trial duration. Participants spent different amount of time for each trial in different response modalities, $F(1, 32) = 77.90$, $p < .001$, $\eta^2 = .73$. Participants spent more time per trial in the slider condition than in the voice condition, $t(30) = 8.71$, $p < .001$, $d = 1.72$, or the self-matching condition, $t(30) = 9.06$, $p < .001$, $d = 1.68$, and spent more time per trial in the self-matching condition than the voice condition, $t(30) = 5.23$, $p < .001$, $d = 0.63$.

Correlations. The correlation between slider and voice performance here did not reach significance, neither for the accuracy, $r(31) = .31$, *ns,* nor error measurements, $r(31) = .35$, *ns*. However, participants with less final error in the self-matching condition did show less final error in the slider condition, $r(31) = .62$, $p < .001$, and in the voice condition, $r(31) = .38$, $p = .037$. The same set of correlations held for the accuracy data, with participants who showed a higher proportion of accurate final responses in the self-matching condition showing a higher proportion of accurate final responses in the slider condition, $r(31) = .65$, $p < .001$, and in the voice condition, $r(31) = .61$, $p < .001$. Participants with higher thresholds for making same/different judgments had more final pitch error in the slider condition, $r(31) = .59$, $p = .001$, and a lower proportion of inaccurate responses in the slider condition, $r(31) = -.50$, $p = .004$, and in the self-matching condition, $r(31) = -.42$, $p = .02$. Same/different judgment thresholds did not correlate with any measurements in the voice condition. Participants with higher thresholds for making higher/lower judgments showed more initial pitch error in the slider condition, $r(31) = .49$, $p = .006$. Finally, to see whether participants made more errors in the self-matching trials to high or low target tones, we correlated each participant's average final pitch error with the height of the target tone. There was no relationship between these two variables, $r(154) = -.11$, *ns.*

After the testing, two of the participants, who were contacted on the basis of their high error rates in this experiment, were identified as probably amusic from their results on the Montreal Battery of Evaluation of Amusia (MBEA; Peretz, Champod, & Hyde, 2003). The analyses were recomputed without the data from these two subjects, and none of the results from the ANOVAs, *t* tests, or correlations changed, with the sole exception that the correlation between same/different judgment threshold and final pitch error in the self-matching condition rose to significance, $r(29) = .45$, $p = .015$. In addition, the mean higher/lower judgment threshold fell from 25 to 20 cents. Other subjects of interest, including some from Experiment 1, were unavailable for further testing.

Discussion

Overall, performance on the slider and voice conditions replicated that of the nonmusicians in Experiment 1. Participants performed much more accurately in the self-matching condition than

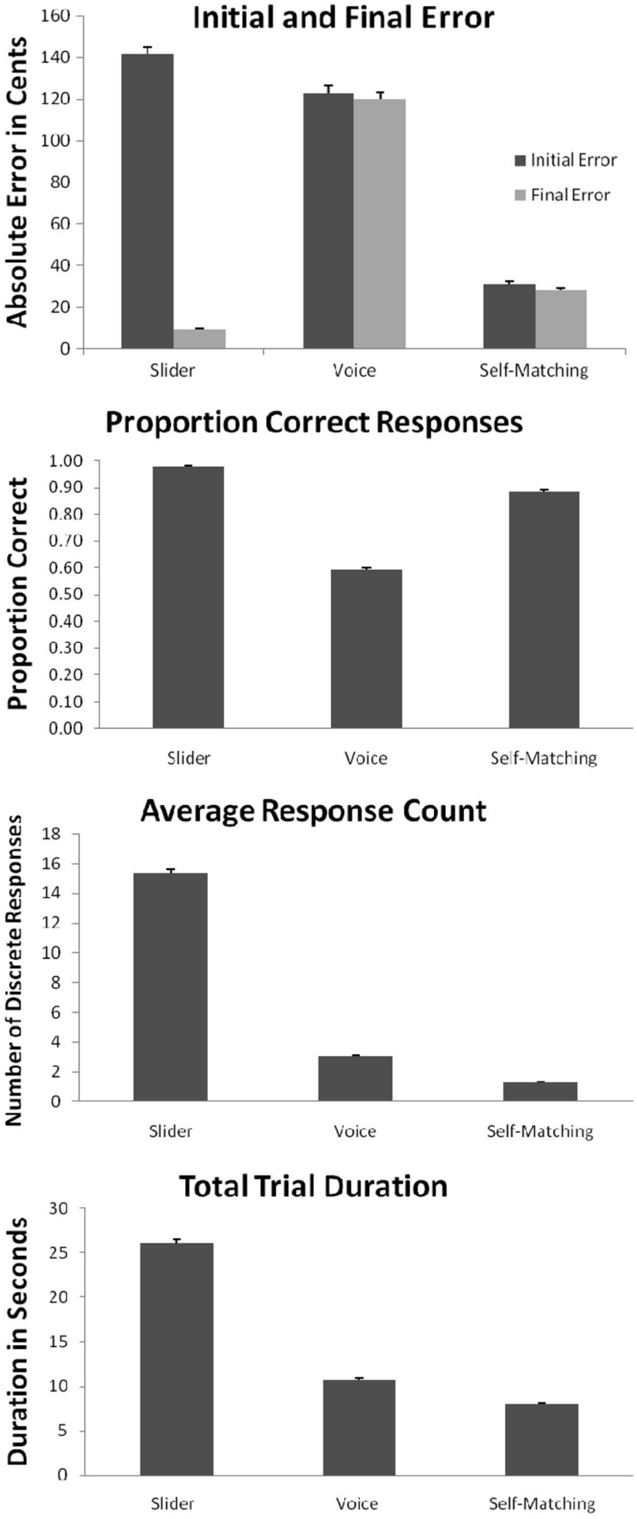

Figure 6. The average absolute value of the pitch error during the initial and final tone produced, the proportion of accurate final responses, the average number of discrete responses, and the total trial duration, shown for participants in the slider, voice, and self-matching conditions from Experiment 2, with standard error bars.

in the voice condition but not as accurately as in the slider condition. This pattern held with both the final response error and the proportion correct measurements. Despite the fairly close match between the slider/target timbre and a typical voice timbre, most participants were aided by the exact match between their own voice and recordings thereof. However, they were still considerably less accurate than they were with the slider. This replicates the findings of Experiment 1, showing that vocal pitch-matching abilities were not limited by perceptual abilities. On the basis of this, it is clear that not all of the singing problems seen in this experiment and Experiment 1 were due to timbral problems. Poor singing seems to have roots in both motor control and timbral difficulties.

Because the stimuli in the self-matching condition were recorded samples, rather than generated tones, it was not possible to play them continuously, as in the voice and slider conditions. However, participants were able to replay the sounds as often as they liked. It should be noted that this discrepancy would tend to work against accurate pitch matching as participants did not generally choose to hear the target tone again after their singing attempt to compare. In addition, the target sounds, being recordings, were not guaranteed to be stable in pitch, as they were in the slider and voice conditions. Neither of these factors seemed to harm participants' ability to match their own recordings.

Participants tended to perform the self-matching task in a way that was very similar to how they performed the pitch-matching task in the voice condition. Participants chose to make fewer responses over a shorter duration of time in the self-matching condition, and the values are much more similar to those of the voice condition than the slider condition. Participants tended to respond only once in each self-matching trial, likely resulting from the slight change in design. However, they were accurate despite the lack of error-correction attempts.

In the self-matching condition, unlike in the voice condition, participants not only matched their own vocal timbre but chose their own target pitches. Figure 5 shows that the targets chosen in the slider/voice conditions (B3 to G4, as measured here) fell within the common range of participants' self-chosen targets. However, participants on average chose target pitches with a lower mean than the targets of the other conditions, and the targets they chose were right-skewed (the median was lower than the mean). This skewness was not an effect of gender or of combining data across participants and seems to reflect a general preference for producing lower tones. This may be due to the fact that low tones require less effort to produce than high tones, or it may be due to singers' misestimations of the midpoint of their own vocal ranges. Because target tones in the self-matching condition did not use the same pitches, or even circumscribe the same intervals, as in the slider and voice conditions, the factor of target pitch could not be included in an ANOVA to compare across conditions. However, due to the lack of effect of target pitch in Experiment 1, this is unlikely to be a complication for the slider and voice conditions, and most likely did not impede their performance in the self-matching condition either, as participants were instructed not to (and could not) produce target pitches out of their own range. Indeed, there was no correlation between average error and target pitch height, as one would expect if there were an effect of target pitch range; therefore, a vocal range explanation of the difference between the self-matching and voice conditions does not seem to hold any weight.

In order to compare perceptual ability as measured by the slider with pitch thresholds as done in prior studies, we measured the correlation between the two. The expected correlation was found—participants with less error on the slider tend to have lower discrimination thresholds. It is interesting that the same correlations were found between discrimination thresholds and performance in the self-matching task. This may reflect participants' greater ability to specify the pitch aiding in finding the vocal match. However, if this were the case, we would also expect pitch discrimination thresholds to correlate with performance in the voice condition, which was not found. Thus, there may be some unknown variable mediating these correlations in this case. The only significant correlation that was found involving up/down judgment thresholds was with initial slider error, which seems to reflect participants' abilities to determine whether the slider target on a new trial is higher or lower than the target on the previous trial and to what extent. These findings support the notion that the slider may tap the same perceptual processes as do traditional pitch perception tasks but do so with more precision (as participants can adjust their own response) and with a much greater wealth of data about the perceptual processes.

Finally, two caveats should be considered when one is interpreting the self-matching data presented here. First, rather than relying on their perception of their own-voice target tones, participants could simply have been remembering how they produced a high tone or a medium-low tone and replicating that. However, such a memory-dependent strategy in a pitch-matching task only would need to be used if the singer were very poor at perceiving the pitch of the tone, which is shown not to be true in most cases by the slider results. In addition, this strategy could be used only in the self-matching condition, not the original voice condition. While the results of this experiment cannot rule out a memory strategy, we consider it unlikely, due to the variance found in poor singers pitch matching (Experiment 1; Pfordresher et al., 2010), the difficulty in remembering the same vocal pitch across the span of time within the experiment, and the fact that singers did not know which of their examples were chosen to appear in the matching portion of the experiment. A strong test of this explanation, however, would be asking singers to match pitch-shifted versions of their self-generated target tones. We would predict that nonmusicians would be just as able to perform this task; however, this has yet to be tested empirically.

The second caveat is that even when participants hear recordings of their own voices, the timbre of the targets is not exactly the same between target and responses in the self-matching condition, due to factors such as bone conduction and the Lombard effect. Everybody thinks they sound different when they hear a recording of themselves, and thus participants' internal representations of their own vocal timbre will not be exactly the same as what they hear from the recordings. However, from the results of the experiment, it seems clear that it is close enough to significantly aid vocal pitch matching compared with the original voice condition. Errors in the original voice condition may be related to the manner in which participants performed this task. Experiment 3 examines whether short-term practice and error correction can aid singers in this voice condition.

Experiment 3

Across both previous experiments, participants had more final error in the voice condition than in the slider condition. However, their initial error is similar in these two conditions. Participants spend more time adjusting their responses and make more attempts to match pitch in the slider condition than in the voice condition and ultimately improve their response over the course of a trial. In order to determine whether the advantage in the slider condition compared with the voice condition is due to the greater number of response attempts in the former, we tested the voice condition again but required participants to make at least 20 different responses in each trial. This surpasses the average number of responses in the slider condition. If participants show improvement over each response within a trial, this would show that they can use error-correcting mechanisms to aid their vocal pitch matching.

Method

Participants. Participants were 11 nonmusicians, all of whom had participated in Experiment 2. Subjects included both good and poor singers, as measured in Experiment 2.

Stimuli and equipment. Stimuli and equipment were the same as in Experiment 1.

Procedure. The procedure was modified from the voice condition from Experiments 1 and 2. In this session, the experimenter controlled the beginning and ending of trials, and participants were instructed to make 21 different pitch-matching responses during each trial. Participants were told to listen to each response they made and to try to be more accurate with each response, if possible. As in the voice condition of the previous experiments, the target tone came back on at any time during a trial when the participant was not singing. The experimenter kept track of the number of responses made by the participant in each trial. Only the first 20 responses were kept and analyzed to guard against counting errors by the experimenter. Due to the length of each trial and the strain on the voice, each subjects was given only 30 trials; the experiment lasted about half an hour.

Results

Figure 7 shows the mean pitch error across the 20 responses. A repeated-measures ANOVA showed no difference between the mean pitch error of the 20 responses, $F(3, 28) = .695$, *ns*, $\eta^2 = .07$. The good singers (as measured by their performance in Experiment 2) had an average error of 24 cents across all trials, and the poor singers had an average error of 147 cents.

Discussion

Even when given multiple opportunities to correct their pitch, participants showed no improvement across the 20 responses within a trial. None of the singers showed any change across their responses—the 20th response tended to be just as prone to error as the first. This shows that the error in the voice condition for Experiments 1 and 2 is likely not due only to the way participants responded and their lack of error-correcting effort. Rather, it seems that participants tended to make few responses in these conditions because they determined that further responses would not aid their accuracy. Participants simply were not able to improve their re-

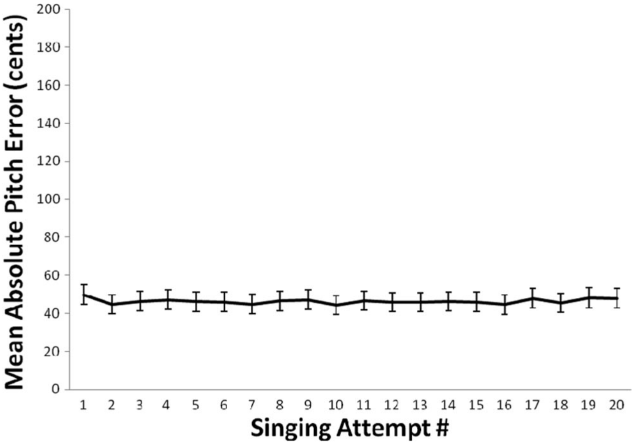

Figure 7. The mean pitch error across the 20 responses in a trial from Experiment 3, with standard error bars.

sponses. It is interesting that even though inaccurate singers produced a consistent inaccurate response with any given trial, across trials, they produced very different inaccurate responses. Average pitch values across the 20 responses could vary by as much as 750 cents between trials with the same target pitch with very little change within the separate responses of a trial. Thus, it seems like their inaccuracy is more an effect of being "locked in" to a particular note within a trial, rather than reflecting a particular mismapping of the stimulus pitch to a vocal output.

Experiment 4

One key difference between slider and singing conditions in these experiments is the visual feedback. Participants may be better able to match a pitch with the slider than their voice because they do not have any visual feedback in the latter condition. To this end, we replicated the voice condition from the previous three experiments, but added a visual representation of the pitch height of the voice as it was produced. Each participant attempted to match target tones both with and without the visual feedback designed to approximate the visual feedback from the slider. If lack of visual feedback was the cause of poorer performance on the voice conditions than the slider conditions, participants should perform better when the visual feedback is present.

Method

Participants. Participants were 16 nonmusicians (12 women and 4 men), none of whom had participated in any of the prior experiments. They had less than 2 years of formal training (M = .42 years) and ranged in age from 19 to 29 (mean age = 23.19 years).

Stimuli and equipment. Stimuli and equipment were the same as in Experiment 1, but a visual component was added to the display. A bar was shown on the computer screen, approximately 50 cm long (about the same length as the physical slider from previous experiments). The Max/MSP program was programmed to calculate the instantaneous frequency of the sung tones and display a graphical representation of that frequency on the bar as it was sung. The visualization bar was given the same maximum and minimum as the slider: 220 Hz to 440 Hz for women and half that for men. Thus, participants had a visual representation of the pitch of their sung response. Note that no information about the visual location of the target was given in the visualization bar. The bar could be turned on or off by the experimenter.

Procedure. The procedure was modified from the voice condition from Experiments 1 and 2. Each participant participated in 50 trials with the visualization bar and 50 trials without the bar. The order of the two halves was counterbalanced across subjects. Participants were given five practice trials before the start of each half of the experiment. They were also given an explanation and demonstration of the visualization bar and were allowed to experiment with it if they wished (by singing alone) before their practice trials with it. The experiment lasted about 25 min on average.

Results

The same six measurements were taken as in Experiment 1, the initial pitch error, final pitch error, proportion of correct responses, response count, total trial duration, and the number of pitch changes in the incorrect direction. To check for differential effects of the visualization between accurate and inaccurate singers, we divided participants into two groups on the basis of their average final pitch accuracies. Good singers (eight total) were those who were within 50 cents of the final pitch on at least 90% of trials and averaged less than 50 cents final pitch error; poor singers (eight total) were those who were within 50 cents of the final pitch on less than 50% of trials and averaged more than 100 cents final pitch error (no participants fell between those two criteria in this study).

Figure 8 shows the mean initial and final pitch error for singers, both with and without the visualization. Six separate 2 × 2 mixed-design ANOVAs (one for each measurement) using the within-subject factor of visualization (present or not) and the between-subjects factor of singing accuracy (good or poor) were conducted. Initial pitch error, final pitch error, and proportion of correct responses all showed a main effect of singing accuracy, which is unsurprising as the participants were divided into groups according to those factors. However, none of these measurements showed a main effect of visualization, nor an interaction between visualization and singing accuracy. In addition, there were no effects of any variables on the number of pitch changes in an incorrect direction. However, visualization did have an effect on

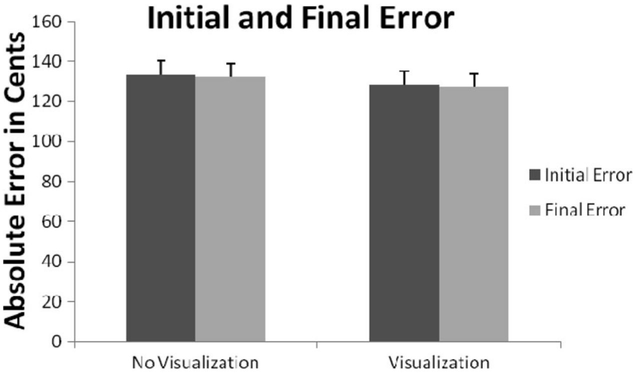

Figure 8. The mean initial and final pitch error for singers with and without the visualization, with standard error bars.

the number of responses, $F(1, 14) = 5.15$, $p = .04$, $\eta^2 = .26$; participants made more responses when they sang with the visualization (2.54) than without (2.31). There was no main effect of or interaction with singing accuracy for this measurement. There was also a main effect of visualization on the total response time, $F(1, 14) = 9.40$, $p = .008$, $\eta^2 = .26$, and an interaction of visualization and singing accuracy, $F(1, 14) = 13.20$, $p = .003$, $\eta^2 = .36$. Good singers took less time on trials with the visualization (7.36 s) than without (7.56 s), but poor singers took considerable longer with the visualization (10.63 s) than without (8.62 s) and longer on average than good singers. There was no main effect of singing accuracy on this variable.

Discussion

Having a visualization of the sung pitch made no difference in any of the measurements of singing accuracy. Neither good nor poor singers were any more accurate when singing with a visual representation of their pitch than they were without it. However, participants did not simply ignore the visualization. They made more responses on average when it was present, and poor singers in particular spent more total time on each trial when singing with the visualization. Singers in general were interested in the visualization, and it seems that poor singers in particular attempted to use this extra source of information to aid their singing accuracy but to no avail. These data suggest that the slider advantage found in this and the previous experiments was not due to the fact that there was only visual feedback in the slider condition but not the singing conditions. Although it is possible that using a visual representation of the target pitch in addition to the sung pitch may aid performance, this would not be analogous to the slider condition, which had no visual indicator of the target pitch. By itself, adding visual feedback to the voice does not improve singing ability, ruling out this possible explanation of poor singing.

Interestingly, no matter the manipulation, singers never seemed to be able to produce tones with average errors below 10 cents, and even good singers tended to have average errors in the 15–30 cent range, even in the self-matching condition. Experiment 5 examines a possible perceptual basis for this finding.

Experiment 5

During the course of the previous experiments, two interesting facts became apparent. First, vocal production was rarely more accurate than 10 cents off from the target—only two musicians in Experiment 1 were more accurate than 8 cents average error—and almost never as accurate as with the slider. Second, none of the experimenters (all trained musicians) could detect the inaccuracies during the testing session, up until approximately 30 cents of error. Indeed, the analyses revealing the inaccuracies were initially surprising, given that the singing had sounded as if it were completely accurate, and 10 or 20 cents of inaccuracy is generally easily detectable to a trained ear (Zwicker & Fastl, 1999). In contrast, similarly sized errors on the slider were readily detectable. Therefore, we hypothesized that it may be more difficult for listeners to perceive tuning discrepancies in the voice than in other instruments. If this is the case, this may explain the lingering small errors in the voice, even among good singers: such errors are not perceivable. Experiment 5 tested participants' ability to perceive tuning discrepancies in the voice compared with the synthesized vocal timbre. We compared the examples of vocal tones produced by participants in Experiment 2 with synthesized versions of the same tones, matched in mean pitch and duration. We hypothesized that listeners would more easily perceive tuning differences in the synthesized voice timbre than in the real voice examples.

Method

Participants. Participants were 28 nonmusicians (24 women and four men) and 15 musicians (12 women and three men), none of whom had participated in prior experiments. Nonmusicians had less than 2 years of formal music training ($M = 0.4$ years), and musicians had more than 6 years ($M = 11.2$ years). Participants ranged in age from 18–30 years ($M = 21.9$ years). No subjects reported any diagnosed hearing deficits or neurological disorders.

Stimuli and equipment. Stimuli were single tones, either sung or sounded with a synthesized vocal timbre. The sung stimuli were taken from the self-produced target tones in the self-matching condition of Experiment 2. All five target tones from 15 of the 31 participants (7 men, 8 women) were selected for use in Experiment 4, making 75 total sung tones. The 15 subjects were selected to be a mix of good, poor, and mediocre singers. The sung tones ranged in mean pitch from slightly higher than F2 (89.5 Hz) to slightly lower than E5 (680.9 Hz), with a mean at around A3 (222.9 Hz). Tones ranged in duration from 2,093 ms to 3,750 ms ($M = 2,871$ ms). Synthesized versions of each of these 75 tones were created, each with the same mean pitch and duration as the original vocal tone. These versions used the same synthesized vocal timbre as the slider as well as the target tones of the slider and voice condition in Experiments 1 and 2 (see Figure 1), including the same 5 Hz FM vibrato.

In order to rule out effects of vibrato on the judgment, we calculated the characteristics of the vibrato for each of the original sung stimulus tones that were judged. The onsets and offsets were cut from each tone, and vibrato data were calculated using an in-house Matlab function which measured (windowed) local maxima and minima of the pitch. Vibrato rate (measured in hertz) was slightly faster in the voice stimuli ($M = 5.58$ Hz, $SD = .73$ Hz) than in the synthesized voice stimuli (set at 5 Hz), $t(73) = 6.80$, $p < .001$, $d = 0.79$. Vibrato amplitude (measured in cents) was larger in the voice stimuli ($M = \pm$ 11.56 cents, $SD = \pm$ 4.16 cents) than in synthesized voice stimuli (set at ± 7 cents), $t(73) = 9.37$, $p < .001$, $d = 1.10$. In addition, we also measured the standard deviation of the pitch of the voice stimuli ($M = 16.12$ cents, $SD = 8.57$ cents), which was slightly higher than the synthesized voice stimuli (set at 8.57 cents), $t(74) = 7.70$, $p < .001$, $d = 0.88$. The synthesized voice was approximately one standard deviation away from the voice mean in each of these measurements.

Both voice and synthesized voice stimuli were normalized for amplitude. Following this, we created 20 different mistuned versions of each of the 75 voice and synthesized voice tones with Melodyne (Celemony Software, Munich, Germany). This mistuned versions ranged from 100 cents flat to 100 cents sharp, in 10-cent intervals. Participants heard the stimuli over the same headphones as in Experiments 1–3, and E-Prime 2.0 (Psychology Software Tools, Sharpsburg, PA) was used to conduct the experiment.

Design and procedure. In the experiment, participants heard a pair of single tones and decided whether the second note was the same as the first tone and in tune, whether it was the same note but out of tune, or whether it was a different note altogether. This categorization method was chosen to help nonmusicians understand the task, and these directions were clearly understood by all participants. Each tone in a pair was taken from the same base note (subjects were never asked to compare tunings across singers or timbres). Ten pseudorandomized lists of 240 tone pairs were created, each containing equal numbers of all variables manipulated, including timbre (voice or synthesized), shift degree, and base note. Half of the tone pairs in each list were exactly the same, and half of the tone pairs were different in pitch. In addition, half of the pairs in each list began with a shifted version, and half begun with an unshifted version. These two manipulations were orthogonal, creating four different possible stimulus orderings for each tone pair. Five practice trials with feedback were presented at the beginning of each list. Participants began the session by filling out a questionnaire on their musical background. The entire experiment lasted about 45 min.

Results

Overall, participants musicians judged 51% of the stimulus pairs as "in-tune," 23% of the pairs as "out-of-tune," and 26% of the pairs as "different note altogether." For nonmusicians, these numbers were 58% "in-tune," 20% "out-of-tune," and 22% "different note altogether." The proportion of hits minus false alarms was calculated for each level of each condition. Judgments of "out-of-tune" and "different note altogether" were both counted equivalently as hits for any stimulus pair with a tuning difference but as false alarms for any exactly equivalent pair of tones. Figure 9 shows the data for musicians and nonmusicians in the voice and synthesized voice conditions. A 2 × (2 × 10) mixed-design ANOVA was conducted with the factors of musical experience (musician and nonmusicians), timbre (voice and synthesized voice), and tuning deviation (the absolute value of the deviation, from 10 to 100 cents) using the hits minus false alarms data. All values reported are Greenhouse–Geisser corrected and so use adjusted degrees of freedom. We showed a main effect of timbre, $F(1, 41) = 106.40$, $p < .001$, $\eta^2 = .09$. All participants were more accurate in discriminating synthesized voice examples compared than actual voice examples. There was also the expected main effect of tuning deviation, $F(5, 184) = 108.90$, $p < .001$, $\eta^2 = .47$, with participants responding more accurately to large tuning differences than to small tuning differences. Timbre interacted with tuning deviation, $F(6, 230) = 8.70$, $p < .001$, $\eta^2 = .03$, such that the discrimination advantage of synthesized voice examples was larger with smaller tuning deviations. We also found a main effect of musical experience, $F(1, 41) = 8.87$, $p = .005$, $\eta^2 = .18$, with musicians performing more accurately than nonmusicians. Musical experience did not interact with timbre, $F(1, 41) = .07$, *ns*, $\eta^2 < .001$, but there was an interaction between tuning deviation and musical experience, $F(5, 184) = 2.44$, $p = .04$, $\eta^2 = .01$, as well as a three-way interaction among musical experience, timbre, and tuning deviation, $F(6, 230) = 3.65$, $p = .002$, $\eta^2 = .01$. Nonmusicians performed considerably worse than musicians on the voice trials compared with the synthesized voice trials; this difference was especially large for judgments of the smaller tuning deviations. To further examine the interaction between instrument and tuning deviation, we conducted a set of follow-up comparisons, with 20 paired-samples *t* tests comparing the hits minus false alarms rate separately for musicians and nonmusicians between the two instruments at each tuning deviation. The alpha was adjusted using the Bonferroni correction for multiple comparisons (adjusted α for 20 comparisons = .0025). These tests showed significant differences for nonmusicians at each tuning deviation between 30 and 70 cents, inclusive, $t(27) > 3.84$ for each t, $p < .0025$, $d > 0.72$. There were significant differences for musicians at each tuning deviation between 20 and 40 cents, inclusive, $t(14) > 3.71$ for each t, $p < .0025$, $d > 0.95$.

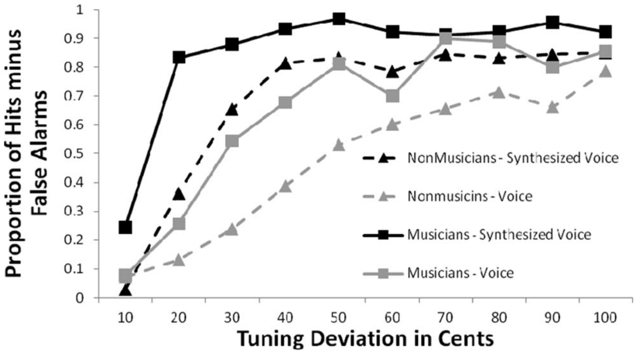

Figure 9. The mean proportion of hits minus false alarms for musicians and nonmusicians comparing two tones with voice and synthesized voice timbres at each pitch difference (in cents) from Experiment 4. Hits here indicates tones of different pitch judged as out-of-tune or a wrong note, false alarms indicates tones of the same pitch judged as out-of-tune or a wrong note. Binomial confidence intervals are omitted here for clarity, but do not exceed ±.08 for nonmusicians and ±.11 for musicians.

We performed two separate ANOVAs to look for stimulus effects (which could not be included with the main ANOVA, due to the fact that the design would be incomplete across the Tuning Deviation variable). First, we conducted a 2 × 2 × 2 mixed-design ANOVA with the factors of musical experience, timbre, and mistuning direction (whether the mistuning shift applied was sharp or flat), using the hits minus false alarms data. There was no effect of mistuning direction, nor any interactions with other variables, confirming that participants were not affected by the direction of the mistuning shift. Second, we performed a 2 × (2 × 2 × 3) mixed-design ANOVA using the factors of musical experience, timbre, source gender (whether the voice or synthesized voice stimulus originated from a female or male singer), and source singing ability (whether the stimulus originated from a good, mediocre, or poor singer). Note that singers here refer to the participants in Experiment 2, from whom these stimuli were derived, and we do not refer to, nor did we measure, the singing ability of the participants in the current experience. For the purposes of this analysis, good singers were classified as those who performed with more than 90% accuracy in both the voice and self-matching conditions in Experiment 2, mediocre singers were classified as those who performed with more than 90% accuracy in the self-matching condition, but less than 90% accuracy in the voice condition, and poor singers were those who performed with less than 90% accuracy in both conditions. There were five good source singers, six mediocre source singers, and four poor source singers, and each category had approximately equal numbers of males and females. Effects of timbre and musical ability are the same as discussed earlier. There was a very small tendency for participants to be more accurate in judging stimuli originating from female sources than from male sources (mean difference = .03 hits—false alarms), but this failed to reach significance, $F(1, 41) = 3.89, p = .056, \eta^2 = .007$. The interactions of source gender and timbre, $F(1, 41) = 3.37, p = .07, \eta^2 = .003$, and source singing ability and timbre, $F(2, 78) = 2.74, p = .07, \eta^2 = .005$ also approached but did not reach significance. There was no main effect of source singing ability, nor any other significant interactions among source singing ability, source gender, and any other factors.

Discussion

The results of this experiment showed a clear effect of timbre on tuning judgments. Notes were more likely to be judged as in tune when they were sounded with the natural voice than with a synthesized vocal timbre, even for exactly equivalent mistuning. On average, nonmusicians showed that they will judge a sung tone as in tune until it is out of tune by at least 50 cents (half a semitone), and at least 30 cents of mistuning was needed for musicians to do the same. To achieve the same benchmark of accuracy (50% accuracy) with the synthesized voice, nonmusicians needed 30 cents of mistuning, and musicians needed only 20 cents. Participants were more accurate in judging mistuning with the synthesized voice timbre for smaller tuning deviations, which likely indicates a ceiling effect in judgments for larger tuning deviations. Musicians were consistently better than nonmusicians, which most likely is a reflection of their greater experience in making tuning judgments, but nonetheless showed an advantage for judging the synthesized voice.

The experimental context is equivalent to a same/different judgment task, but it may be that participants' judgments would be different if we had asked them to explicitly make a same/different judgment rather than an in-tune/out-of-tune judgment. However, Warrier and Zatorre (2002) showed that there are no differences between the behavior of participants making these two types of judgments; thus, we do not have any reason to presume that this had any effect on the results. Neither musicians nor nonmusicians showed any confusion about how to make tuning judgments and used both "same note but out-of-tune" and "wrong note altogether" responses in appropriate ways. The use of two types of "different" responses was intended to prevent listeners from judging two tones as the same if they perceived the second as an example of a variant on the same type of note, and subjects did not indicate that they found this distinction confusing in any way. Peretz, Brattico, Järvenpää, and Tervaniemi (2009) showed different neural and behavioral responses to mistuned and out-of-key tones, also confirming the distinction made by participants in the current study.

These findings confirm prior studies showing the interrelationship of timbre and pitch (e.g., Krumhansl & Iverson, 1992; Melara & Marks, 1990a, 1990b, 1990c; Pitt, 1994; Warrier & Zatorre, 2002). Note that in the present experiment, the timbre was always consistent throughout a trial (i.e., violin was never compared with voice), and we compared how stably pitch is represented within two different timbres. The more stably a pitch is represented within a timbre, the smaller the range of acceptable tuning should be for tones of that timbre. This concept is related to but may not be identical to that of pitch salience or clarity within a timbre. The differences we found between the voice and the synthesized voice timbre showed a different kind of interaction between pitch and timbre, and point to the importance of considering timbre in pitch perception or production tasks.

Our acoustic analyses of the stimuli showed that the voice data were slightly more variable in pitch and included slightly more vibrato than the synthesized voice tones. Vibrato is an interesting quality, standing on the cusp of pitch and timbre. While it is defined as a more-or-less sinusoidal variation of the pitch of a tone, it is generally not perceived as a fluctuation of pitch. Rather it is generally perceived as a timbral quality of a steady pitch, when not used to excess and is often thought of as adding "color" to a tone. The pitch of tone performed with vibrato is perceived as the mean of the fluctuating pitch, even for very large vibrato amplitudes (Shonle & Horan, 1980; Sundberg, 1978). However, adding more variation in pitch may serve to make the representation of that pitch less certain. Yoo, Sullivan, Moore, and Fujinaga (1998) noted that many musicians believe that vibrato is an effective way of hiding tuning errors, and van Besouw, Brereton, and Howard (2008) showed that synthesized tones with vibrato had a greater range of acceptable tuning (as measured by trained musicians) than those without vibrato, by about 10 cents. Our synthesized stimuli included slightly less vibrato and pitch variation than the original voice stimuli, but were only approximately 1 standard deviation from the mean of the original voice stimuli. Because our synthesized stimuli would not be considered as an outlier among the actual stimuli, it is unlikely that vibrato and pitch variation can account for the large tuning acceptability differences between the two timbres.

In sum, these data show that vocal tones are perceived less accurately in pitch than synthesized voice tones, and that listeners are more forgiving of tuning errors in vocal tones than in synthesized voice tones. It is possible that the difference between the two timbres arises partially from acoustic differences, including vibrato or variation, or perhaps some other unmeasured variable, but it is also possible that top-down factors may account for the difference. Listeners may have implicitly learned to disregard tuning errors when they appear in the voice, and it is conceivable that this represents an inborn, automatic feature of voice processing. This lower ability to perceive vocal mistuning also carries over when compared against a natural instrument and in a melodic context as well (Hutchins, Roquet, & Peretz, in preparation), and so is unlikely to be specific to these particular stimuli or the experimental context. This difference provides a possible explanation for the persistent 10–30 cent pitch errors seen in the good singers in the previous three experiments: Listeners simply do not reliably hear such small pitch errors as out of tune. These results also help to confirm our earlier use of 50 cents as a cutoff criterion between good and poor singers, because this is the point at which most listeners begin to hear a sung tone as out of tune.

General Discussion

In the studies presented here, we have introduced a new method of studying pitch perception and production. The slider allows us to study these issues using an active response that is similar in many important ways to the voice. Prior studies (e.g., Bradshaw & McHenry, 2005; Dalla Bella et al., 2007; Pfordresher & Brown, 2007; Watts et al., 2005) have used pitch discrimination tasks as a comparison to vocal production tasks. Because of the differences in these two tasks, they cannot be compared directly but only through a correlation. This makes it difficult to speak directly to the influence of pitch perception on production abilities. The slider, in contrast, does provide a pitch matching task that can be directly compared with vocal pitch matching. This is not only because it provides measurements in the same units but also because it bears many important similarities to the vocal mechanism, including a linear dimension of pitch that can produce only one pitch at a time and can easily be in an on-or-off state. Experiments 1 and 2 here show conclusively that the large majority of nonmusicians are able to use the slider reliably and accurately, with a precision within 10 cents, even when they were inaccurate in the vocal task. Experiment 2 showed that slider-matching abilities correlate with standard measurements of pitch discrimination and in fact can be more sensitive to perceptual abilities than discrimination tasks. Thus, the slider is measuring the same type of ability as investigated in prior studies but in a metric that supports direct comparison to vocal pitch matching abilities. The slider also proved to be a potentially useful diagnostic task, as two participants who performed poorly on the slider ended up being diagnosed as amusic.

In vocal pitch matching, Experiments 1 and 2 showed that singing abilities could vary widely across individuals. Musicians were more accurate in their singing than nonmusicians, but many nonmusicians were quite accurate despite their lack of musical training. Across experiments, 20 of 53 nonmusicians (38%) correctly sang the synthesized target vocal tone with their voice more than 90% of the time, whereas 25 of 53 nonmusicians (47%) failed to match the tone on even 50% of the time. We used the criterion of 90 % accuracy (10% error rate) here to divide good singers from poor singers, with accurate responses being those less than 50 cents from the target tone. Singers classified as good singers, on the basis of having at least a 90% accuracy rate, tend to have mean errors of 50 cents or less, but we used the accuracy rate as a criterion here because it makes little difference whether a tone is off by, for example, 200 cents or 600 cents: They are both errors and equally wrong. Furthermore, as we saw in Experiment 5, listeners have a hard time distinguishing between perfect matches and errors less than 50 cents.

Our finding of 62% of nonmusicians as poor singers is a much higher incidence of poor singing than has been reported in prior studies (typically around 10%–20%, e.g., Dalla Bella et al., 2007; Pfordresher & Brown, 2007). Partially, this is due to definitional differences, as those studies used a cutoff criterion of 100 cents to distinguish accurate from inaccurate singing, whereas we used a criterion of only 50 cents, which was confirmed as a reasonable one by Experiment 5. However, even if we use the more liberal definition of poor singers as those with average errors over 100 cents, as used by Pfordresher and Brown (2007), we still find 21 of our 53 nonmusicians (40%) defined as poor singers. The differences between this study and prior work may be a function of the type and complexity of the stimuli used, or perhaps the precise timbre of the target tones (which can have a strong effect, as we see from Experiments 2 and 5). Whatever the cause, this study shows that under certain conditions, poor singing may be more prevalent than prior studies had estimated. In fact, a later inventory of nonmusicians, reported in Pfordresher and Mantell (2009), showed a higher incidence of poor singing than earlier studies as well, with 24% showing mean errors over 100 cents and more than 40% showing mean errors over 50 cents, bringing their estimates more in line with those found in our studies. Nonmusicians' low estimations of their singing abilities may have some grounding in truth.

Experiment 2 showed, however, that even poor singers were much improved when their own voices were used as target tones. Despite the fact that the synthesized voice tones were very similar in timbre to an actual voice, it seems that the subtle timbral difference between that and the participants' own voice is responsible for a large change in performance. Twenty-three of the 31 nonmusicians (74%) who participated in Experiment 2 matched their own target pitches 90% of the time or more, and only three (10%) failed to match the pitch 50% of the time or more. This improvement cannot simply be a result of participants producing target pitches in a more comfortable range, since there were no target-pitch related effects in Experiment 1, and there was no correlation between average pitch errors and target pitch heights in Experiment 2. The general benefit of hearing one's own voice as a target may be due to the exactness of the timbral match but could also include an effect of familiarity with the sound of one's own voice. The self-matching benefit may also arise not simply from a familiarity with perceiving one's own voice but with a familiarity of producing and controlling one's own voice. It is possible that hearing one's own voice evokes the feedback control loop of the vocal system (Burnett & Larson, 2002), although in our case, it is clear to the participant that the target note is not currently being produced by himself or herself. Finally, the benefit of matching one's own voice also suggests possible treatments or pedagogical approaches to improve singing among poor pitch matchers, such as recording and editing samples of participants' singing for use as examples.

Both Experiments 1 and 2 showed differences in the way participants performed the slider and voice matching tasks. Whereas participants would spend a lot of time correcting and perfecting their responses with the slider, they would not do so with their vocal matching, despite being given the opportunity. Experiment 3 showed that this difference in approach to the two tasks was not responsible for the better performance on the slider. Even when participants were forced to make multiple responses and explicitly instructed to improve their response each time, they failed to do so. Neither good nor poor singers made any substantial adjustments to their pitch over the 20 responses in a trial. People seem to perform pitch matching differently when using their voice, compared with an instrument, in a way that is consistent across individuals. In fact, some subjects indicated that they understood that they had not matched the target pitch well, yet persisted to produce the same pitch across the whole trial. These differences in task approach between singing and slider held across all participants, even though the tasks were introduced in the same way to the participants, and care was taken to stress their similarities. This could point to different mechanisms underlying singing and other forms of pitch production, and possibly even different mechanisms for the perception of pitch underlying those two tasks.

One crucial difference between using the slider and using the voice is that while visual feedback is present for the participant using the slider (who can see the position of his or her own hand on the slider), there is no equivalent visual feedback when producing a pitch with the voice. Experiment 4 established that the slider advantage was not due simply to this difference in visual feedback across conditions. Singers were no better able to match a pitch when given visual feedback about the pitch they were currently producing than without such visual feedback. Although the possibility remains that visual feedback may be helpful in other circumstances, it is not at the heart of the difference between slider and singing conditions in these studies.

Experiment 5 demonstrated that the pitch of vocal tones is heard with less accuracy than the pitch of tones played with the slider and provided justification for our use of 50 cents as a dividing line between good and poor pitch matches. Nonmusicians did not recognize a vocal note as out of tune until it was mistuned by over 50 cents. It is likely that the 10–30 cent errors among good singers in Experiments 1-3 happen because both singers and listeners cannot tell that they are in fact mistuned. The 50-cent pitch change detection threshold was much higher than the average error in the slider task. This points to a role of task in measuring perceptual ability and suggests that a direct comparison task such as that used in Experiment 5 and many other perceptual studies may not fully capture the limits of perceptual discrimination ability. Matching paradigms such as the slider may give better estimates of discrimination ability.

Even though people were less accurate at judging mistuning in the voice than the synthesized voice, singers were nonetheless more accurate at matching their own voice than the synthesized voice, which again points to different mechanisms for pitch perception in production and perception tasks. Previous work has suggested that separate neural pathways may underlie perception and production of pitch (Dalla Bella et al., 2009; Hafke, 2008; Hutchins et al., 2010; Loui, Alsop, & Schlaug, 2009; Loui, Guenther, Mathys, & Schlaug, 2008), and these two functions are thought to be instantiated in the superior and inferior routes of the arcuate fasciculus, connecting the auditory cortex to prefrontal regions (Loui et al., 2009). The performance differences between vocal and nonvocal tasks that we see across the five experiments supports this dual-route model of pitch processing.

Across Experiments 1–3, there seems to be a small trend for slightly improving average results in the (synthesized) voice condition; we believe this may be due to a small participant self-selection effect. Participants who dislike singing are less likely to participate in an experiment involving singing, and those same participants are also more likely to be poor singers. Once the experimenters had been testing participants for a while, poor singers may have been less likely to participate than good singers. Furthermore, Experiments 2 and 3 involved multiple sessions, and people who disliked singing were probably less inclined to return for the follow-up sessions. In light of this, it is all the more surprising that we found higher rates of poor singing than previous studies. The final section here returns to the original question of what causes poor singing and focuses on the 31 participants from Experiment 2 who completed the slider, voice, and self-matching subtests to illustrate the multiple causes of poor singing.

What Causes Poor Singing?

In this section, we examine those nonmusicians who participated in Experiment 2. We divided these participants into subsets on the basis of their performance in the slider, voice, and self-matching conditions. Participants were considered to be impaired in a particular condition if they made more than 10% erroneous responses (over 50 cents from the target), which generally corresponds to an average error of approximately 50 cents as well. Although there are eight possible combinations of good and poor performance across the three tasks, only four combinations were actually found among our pool of participants. Three of these four groups correspond to three possible explanations for poor singing performance (the final group was good at each of the three tasks). Although the proportion of participants who fall into each group is determined by the precise criteria used to divide between good and poor task performance, note that while different criteria would create slightly different rates of participants in each subset, the same basic subsets would still be observed.

Perception. Of the 31 participants in Experiment 2, two (6%) showed consistent impairments in their abilities to match pitch on the slider and with their voices to both synthesized and self-produced targets.[1] These participants do seem to be impaired by their abilities to perceive the pitch of tones accurately, and it is likely that their vocal pitch matching problems are the result of their impaired perception. We later identified both of these subjects as probably congenital amusic using the MBEA online test (Peretz et al., 2008). Congenital amusia is identified as a problem involving poor pitch perception and usually associated with poor singing, so this is consistent with the results from the slider. In Amir et al. (2003) and Watts et al. (2005) and in various articles by Moore and colleagues, poor perceptual ability has been argued to be a major cause of poor singing ability. The current set of experiments shows that poor perception can be a cause of poor singing ability but relatively rarely. Furthermore, we found that participants were more accurate at perceiving the pitch of synthesized voice tones than actual voices. If perception errors were the major cause of poor singing, we would expect that participants would be more accurate in matching a synthesized voice timbre than an actual voice timbre. However, we saw that participants were almost always more accurate in matching the pitch of their own voice than the synthesized voice timbre, which further argues against a perceptual account of poor singing. A follow-up test comparing the same/different judgment thresholds across the four groups of participants revealed no significant effect, $F(3, 30) = 2.11$, $p = .12$, $\eta^2 = .19$, further confirming that singing ability differences were not caused by differences in pitch perception ability. Although there should be differences in this measurement between the poor slider performers and the other subjects, these were likely not found due to the small number of poor slider performers (only two); this group in fact showed the highest mean threshold value of the four groups.

We may also rule out the explanation of poor singing being due to decreased auditory sensitivity during vocalization. It is a possibility that hearing levels are attenuated during vocalization and that the listener is not adequately able to judge the pitch of his or her own voice; however, Shearer (1978) showed that less auditory masking is observed from one's own voice than from recordings of one's own voice. The pitch–shift response, where phonating participants will adjust their fundamental frequency in response to an upwards or downwards shift in the pitch of their own vocal feedback (e.g., Burnett, Freedland, Larson, & Hain, 1998; Burnett & Larson, 2002; Natke, Donath, & Kalveram, 2003), also provides

[1] One other participant showed similar errors on the slider task but was omitted from this category because all errors occurred on a particular target pitch, a result not seen in any other participants.

indication that singers can hear the pitch of their own voices while singing and make very fine (unconscious) judgments of their own vocal pitch. Because of this, it is unlikely that adequate ability to hear one's own vocalizations is a cause of poor singing.

Vocal–motor control. Six of the 31 participants (19%) in Experiment 2 had no trouble matching pitches with the slider but could not consistently match pitches with their voice to synthesized or self-produced targets. These participants had no perceptual problems (as measured by their slider performance and pitch discrimination thresholds) but had vocal pitch matching problems with all types of targets nonetheless. These results implicate a motor control problem of the voice as the primary cause of their poor singing. Although few studies have explicitly espoused an effect of poor vocal–motor control or coordination, this is likely because the tools to examine this hypothesis have been lacking, and perceptual problems are more squarely under the streetlamp, so to speak. However, with the slider, which uses a different motor mechanism and set of effectors than the voice, we can see motor control effects. When these participants perform the same task using a different motor mechanism (here, the hand and finger in place of the larynx and vocal folds), their ability to match pitch is much better. This difference in abilities clearly points to participants' problems in controlling the vocal mechanisms that are responsible for the pitch of their voice. Pfordresher and Brown (2007) had mentioned and ruled out a motor control hypothesis, but they based this on lack of effect of vocal range or pitch stability. However, coordination is also a type of motor control problem, and this is revealed by the variance in responses. We found higher variances in poor singers (as did Pfordresher et al., 2010), which is indicative of a motor control problem.

Sensorimotor effects. Pfordresher and Brown (2007) defined a sensorimotor control problem as an imitative deficit, such that singers mismapped pitches onto motor gestures. This would seem to predict that singers with this problem would sing a wrong pitch with low variance and high consistency across attempts. Our data show that this pattern does not hold for the poor singers tested here. These singers showed increasing variance in their singing accuracies associated with higher error rates; those singers who make errors do not do so in a consistent fashion. Across trials, they do not consistently map one input pitch to one output motor gesture. This is actually in agreement with the data presented by Pfordresher and Brown (2007), who showed greater variability in their poor singers, and is expanded upon at length in Pfordresher et al. (2010). Because of this, we do not see evidence for a mismapping problem.

In Experiment 2, however, 11 of 31 nonmusicians (35%) performed well on the slider task and on the self-matching task but did not match the synthesized vocal tones consistently. These singers have no perceptual problems, as evidenced by their slider performance, and their accurate performance on the self-matching task showed that they have no motor control problems. The small difference in timbre between their own recorded voice and the synthesized voice was enough to impair their pitch matching abilities, though. This is evidence that these singers have a problem "translating" one timbre to another. These singers may not have a unitary concept of pitch and could be representing the pitch of different timbres along different dimensions. Although they can determine higher and lower pitches within each timbre, they have difficulty specifying equivalent pitches across timbres. This type of error has often been reported in pitch discrimination tasks across timbres (Pitt, 1994; Pitt & Crowder, 1992; Platt & Racine, 1985; Warrier & Zatorre, 2002). The current experiments show that this is a likely primary cause of these participants' failures to match pitch accurately.

This timbre translation problem is different from Pfordresher and Brown's mismapping hypothesis because it does not specify that there exists a consistent mapping between particular inputs and motor gestures. Rather, there is a lack of a consistent mapping or translation between timbres, which leads to inconsistent pitch productions. This also explains why participants will rarely change their initial responses when vocally matching pitch, even for inaccurate responses. The singer's own vocal response is perceived along a different dimension of pitch from the target tone. Due to this lack of timbre translation, the singer's own inaccurate response is not perceived as higher or lower than the target pitch but simply as different, which impairs the singer's ability to make any substantial improvements to the response, even if he or she can hear that the response is incorrect. However, once other trials have passed and the same target tone returns, the poor singer has lost the memory of that particular (erroneous) response and makes another attempt, often finding a different incorrect pitch. It should also be noted that these types of problems could co-occur with motor control problems, but the latter would tend to mask the effects of the former.

We would hypothesize that singers with a timbre-translation problem would be more able to sing along with a chorus than with a piano, for example, and should have few problems singing back a song from (long-term) memory, as there are fewer timbre translations to be made in those settings. However, they would find problems in a more formal music setting, such as following along with a music teacher playing the piano, which is a common teaching method in schools or formal choirs. In some studies, the role of singing with or without a model has been addressed, with varying results (Hutchins et al., 2010; Pfordresher & Brown, 2007; Tremblay-Champoux et al., in press; Wise & Sloboda, 2008), but none of these have directly investigated the role of the timbre of the model. Our results predict that this factor should affect singing abilities but only among a subset of the population. In contrast with singers with poor vocal–motor control, who sing poorly in most contexts, singers with a timbre-translation problem would seem to have different levels of singing ability when examined in different contexts.

Memory and motivation. The current experiments did not attempt to determine the role of memory or motivation on pitch matching. We considered only problems that can affect the immediate mapping of pitch input to motor output and controlled for memory and motivation factors as well as we could. It seems likely that these issues could affect singing in some settings and contexts (Dalla Bella et al., 2007; Estis et al., 2009, 2010); however, they are probably not the cause of the cases of poor singing reported in the current studies. The targets in both voice and slider conditions were played back continuously and were represented immediately after the participant finished an instance of a production. This allows productions to be immediately compared with the target tone, with no chance for errors to arise in the pitch memory. Productions can also be verified after the fact, as the target tone is represented. Thus, it is very unlikely that a poor representation of the pitch in memory is responsible for instances of poor singing.

The current study attempts to control for some motivational factors through the experimental context, which is in a formal laboratory setting. This was shown to elicit more accurate singing than an informal context by Dalla Bella et al. (2007). An experimenter was present with the subject at all times to verify that each participant was focusing on the task at hand. In addition, these motivational factors should be the same between the slider and voice conditions, as they were performed consecutively in the same locations. The fact that participants in Experiments 2 and 3 returned for multiple sessions also indicates that they were motivated to complete the tasks. Long-term motivational factors cannot be ruled out, of course, as a cause of poor singing ability; however, these do not address the proximal causes of poor singing.

Conclusion

This study documented pitch matching abilities of musicians and nonmusicians, using both a simple instrument (the slider) and the voice. We showed that both groups of participants were considerably more accurate at matching pitches with the slider than with their voice, despite their unfamiliarity with the instrument. As expected, musicians outperformed nonmusicians at both tasks, although many nonmusicians achieved musician-like standards. There was a marked improvement in pitch matching ability for matching one's own voice compared with a synthesized voice tone. All participants approached the slider task differently than they approached the singing task. Small degrees of vocal pitch matching error seem not be detectable, compared with other timbres, and this difference seems to underlie the differences between slider and voice performance that exists among good singers. The findings demonstrate a range of singing abilities and implicate different causes for different subsets of poor pitch matchers. Perceptual deficits are relatively rare but can be a cause of poor singing in some cases. More common are motor control problems and timbre-translation problems, which seem to be the primary causes of poor singing in most cases.

References

Amir, O., Amir, N., & Kishon-Rabin, L. (2003). The effect of superior auditory skills on vocal accuracy. *Journal of the Acoustical Society of America, 113,* 1102–1108. doi:10.1121/1.1536632

Ayotte, J., Peretz, I., & Hyde, K. (2002). Congenital amusia: A group study of adults afflicted with a music-specific disorder. *Brain, 125,* 238–251. doi:10.1093/brain/awf028

Bradshaw, E., & McHenry, M. A. (2005). Pitch discrimination and pitch matching abilities of adults who sing inaccurately. *Journal of Voice, 19,* 431–439. doi:10.1016/j.jvoice.2004.07.010

Burnett, T. A., Freedland, M. B., Larson, C. R., & Hain, T. C. (1998). Voice *F*0 responses to manipulations in pitch feedback. *Journal of the Acoustical Society of America, 103,* 3153–3161. doi:10.1121/1.423073

Burnett, T. A., & Larson, C. R. (2002). Early pitch shift response is active in both steady and dynamic voice pitch control. *Journal of the Acoustical Society of America, 112,* 1058–1063. doi:10.1121/1.1487844

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. doi:10.1121/1.2427111

Dalla Bella, S., Giguère, J.-F., & Peretz, I. (2009). Singing in congenital amusia: An acoustical approach. *Journal of the Acoustical Society of America, 126,* 414–424. doi:10.1121/1.3132504

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. doi:10.1121/1.1458024

Demorest, S. M. (2001). Pitch-matching performance of junior high boys: A comparison of perception and production. *Bulletin of the Council for Research in Music Education, 151,* 63–70.

Demorest, S. M., & Clements, A. (2007). Factors influencing the pitch-matching of junior high boys. *Journal of Research in Music Education, 55,* 190–203. doi:10.1177/002242940705500302

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. doi:10.1016/j.jvoice.2007.10.001

Estis, J. M., Dean-Claytor, A., Moore, R. E., & Rowell, T. L. (2010). Pitch-matching accuracy in trained singers and untrained individuals: The impact of musical interference and noise. *Journal of Voice, 25,* 173–180. doi:10.1016/j.jvoice.2009.10.010

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

Garner, W. (1974). *The processing of information and structure.* Potomac, MD: Erlbaum.

Goetze, M., Cooper, N., & Brown, C. J. (1990). Recent research on singing in the general music classroom. *Bulletin of the Council for Research in Music Education, 104,* 16–37.

Gould, A. O. (1969). Developing specialized programs for singing in the elementary school. *Bulletin of the Council for Research in Music Education, 17,* 9–22.

Grassi, M., & Soranzo, A. (2009). MLP: A MATLAB toolbox for rapid and reliable auditory threshold estimation. *Behavior Research Methods, 41,* 20–28. doi:10.3758/BRM.41.1.20

Green, D. M. (1990). Stimulus selection in adaptive psychophysical procedures. *Journal of the Acoustical Society of America, 87,* 2662–2674. doi:10.1121/1.399058

Green, D. M. (1993). A maximum-likelihood method for estimating thresholds in a yes–no task. *Journal of the Acoustical Society of America, 93,* 2096–2105. doi:10.1121/1.406696

Hafke, H. Z. (2008). Nonconscious control of fundamental voice frequency. *Journal of the Acoustical Society of America, 123,* 273–278. doi:10.1121/1.2817357

Hutchins, S., Roquet, C., & Peretz, I. (Manuscript in preparation). *Perceiving tuning errors in the voice and violin.*

Hutchins, S., Zarate, J. M., Zatorre, R. J., & Peretz, I. (2010). An acoustical study of vocal pitch matching in congenital amusia. *Journal of the Acoustical Society of America, 127,* 504–512. doi:10.1121/1.3270391

Hyde, K. L., & Peretz, I. (2004). Brains that are out of tune but in time. *Psychological Science, 15,* 356–360. doi:10.1111/j.0956-7976.2004.00683.x

Joyner, D. R. (1969). The monotone problem. *Journal of Research in Music Education, 17,* 115–124. doi:10.2307/3344198

Krumhansl, C. L., & Iverson, P. (1992). Perceptual interactions between musical pitch and timbre. *Journal of Experimental Psychology: Human Perception and Performance, 18,* 739–751. doi:10.1037/0096-1523.18.3.739

Loui, P., Alsop, D., & Schlaug, G. (2009). Tone deafness: A new disconnection syndrome? *Journal of Neuroscience, 29,* 10215–10220. doi:10.1523/JNEUROSCI.1701-09.2009

Loui, P., Guenther, F., Mathys, C., & Schlaug, G. (2008). Action–perception mismatch in tone-deafness. *Current Biology, 18,* R331–R332. doi:10.1016/j.cub.2008.02.045

Melara, R. D., & Marks, L. E. (1990a). HARD and SOFT interacting dimensions: Differential effects of dual context on classification. *Attention, Perception, & Psychophysics, 47,* 307–325. doi:10.3758/BF03210870

Melara, R. D., & Marks, L. E. (1990b). Interaction among auditory dimensions: Timbre, pitch, and loudness. *Attention, Perception, & Psychophysics, 48,* 169–178. doi:10.3758/BF03207084

Melara, R. D., & Marks, L. E. (1990c). Perceptual primacy of dimensions: Support for a model of dimensional interaction. *Journal of Experimental Psychology: Human Perception and Performance, 16,* 398–414. doi: 10.1037/0096-1523.16.2.398

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. doi:10.1016/j.jvoice.2006.10.013

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. doi:10.1016/j.jvoice.2006.04.004

Murry, T. (1990). Pitch-matching accuracy in singers and non-singers. *Journal of Voice, 4,* 317–321. doi:10.1016/S0892-1997(05)80048-7

Natke, U., Donath, T. M., & Kalveram, K. T. (2003). Control of voice fundamental frequency in speaking versus singing. *Journal of the Acoustical Society of America, 113,* 1587–1593. doi:10.1121/1.1543928

Nikjeh, D. A., Lister, J. J., & Frisch, S. A. (2009). The relationship between pitch discrimination and vocal production: Comparison of vocal and instrumental musicians. *Journal of the Acoustical Society of America, 125,* 328–338. doi:10.1121/1.3021309

Peretz, I., Brattico, E., Järvenpää, M., & Tervaniemi, M. (2009). The amusic brain: In tune, out of key, and unaware, *Brain, 132,* 1277–1286. doi:10.1093/brain/awp055

Peretz, I., Champod, A. S., & Hyde, K. (2003). Varieties of musical disorders: The Montreal Battery of Evaluation of Amusia. *Annals of the New York Academy of Sciences, 999,* 58–75. doi:10.1196/annals.1284.006

Peretz, I., Gosselin, N., Tillmann, B., Cuddy, L. L., Gagnon, B., Trimmer, C. G., . . . Bouchard, B. (2008). On-line identification of congenital amusia. *Music Perception, 25,* 331–343. doi:10.1525/mp.2008.25.4.331

Pfordresher, P. Q., & Brown, S. (2007). Poor-pitch singing in the absence of "tone deafness." *Music Perception, 25,* 95–115. doi:10.1525/mp.2007.25.2.95

Pfordresher, P. Q., & Brown, S. (2009). Enhanced production and perception of musical pitch in tone language speakers. *Attention, Perception, & Psychophysics, 71,* 1385–1398. doi:10.3758/APP.71.6.1385

Pfordresher, P. Q., Brown, S., Meier, K. M., Belyk, M., & Liotti, M. (2010). Imprecise singing is widespread. *Journal of the Acoustical Society of America, 128,* 2182–2190. doi:10.1121/1.3478782

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). Jyväskylä, Finland: Author.

Pitt, M. A. (1994). Perception of pitch and timbre by musically trained and untrained listeners. *Journal of Experimental Psychology: Human Perception and Performance, 20,* 976–986. doi:10.1037/0096-1523.20.5.976

Pitt, M. A., & Crowder, R. G. (1992). The role of spectral and dynamic cues in imagery for musical timbre. *Journal of Experimental Psychology: Human Perception and Performance, 18,* 728–738. doi:10.1037/0096-1523.18.3.728

Platt, J. R., & Racine, R. J. (1985). Effect of frequency, timbre, experience, and feedback on musical tuning skills. *Attention, Perception, & Psychophysics, 38,* 543–553. doi:10.3758/BF03207064

Semal, C., & Demany, L. (1991). Dissociation of pitch from timbre in auditory short-term memory. *Journal of the Acoustical Society of America, 89,* 2404–2410. doi:10.1121/1.400928

Semal, C., & Demany, L. (1993). Further evidence for an autonomous processing of pitch in auditory short-term memory. *Journal of the Acoustical Society of America, 94,* 1315–1322. doi:10.1121/1.408159

Shearer, W. M. (1978). Self-masking effects from live and recorded vowels. *Journal of Auditory Research, 18,* 213–219.

Shonle, J. I., & Horan, K. E. (1980). The pitch of vibrato tones. *Journal of the Acoustical Society of America, 67,* 246–252. doi:10.1121/1.383733

Sundberg, J. (1978). Effects of the vibrato and the "singing formant" on pitch. *Journal of Research in Singing, 5,* 5–17.

Sundberg, J. (1987). *The science of the singing voice.* Dekalb, IL: Northern Illinois University Press.

Tremblay-Champoux, A., Dalla Bella, S., Phillips-Silver, J., Lebrun, M.-A. & Peretz, I. (in press). Singing proficiency in congenital amusia: Imitation helps. *Cognitive Neuropsychology.*

van Besouw, R. M., Brereton, J. S., & Howard, D. M. (2008). Range of tuning for tones with and without vibrato. *Music Perception, 26,* 145–156. doi:10.1525/mp.2008.26.2.145

Warrier, C. M., & Zatorre, R. J. (2002). Influence of tonal context and timbral variation on perception of pitch. *Perception & Psychophysics, 64,* 198–207. doi:10.3758/BF03195786

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. doi:10.1067/S0892-1997(03)00068-7

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. doi: 10.1016/j.jvoice.2004.09.001

Watts, C., Murphy, J., & Barnes-Burroughs, K. (2003). Pitch matching accuracy of trained singers, untrained subjects with talented singing voices, and untrained subjects with nontalented singing voices in conditions of varying feedback. *Journal of Voice, 17,* 185–194. doi:10.1016/S0892-1997(03)00023-7

Watts, C. R., & Hall, M. D. (2008). Timbral influences on vocal pitch-matching accuracy. *Logopedics Phoniatrics Vocology, 33,* 74–82. doi: 10.1080/14015430802028434

Wise, K. J., & Sloboda, J. A. (2008). Establishing an empirical profile of self-defined "tone deafness": Perception, singing performance, and self-assessment. *Music Scientiae, 12,* 3–26. doi:10.1177/102986490801200102

Yoo, L., Sullivan, D. S., Jr., Moore, S., & Fujinaga, I. (1998). The effect of vibrato on the response time in determining the pitch relationship of violin tones. In S. W. Yi (Ed.), *Proceedings of the 5th International Conference on Music Perception and Cognition* (pp. 209–211). Seoul, Korea: Seoul National University.

Zarate, J. M., Delhommeau, K., Wood, S., & Zatorre, R. J. (2010). Vocal accuracy and neural plasticity following micromelody–discrimination training. *PLoS ONE, 5,* e11181. doi:10.1371/journal.pone.0011181

Zwicker, E., & Fastl, H. (1999). *Psychoacoustics: Facts and models.* Berlin, Germany: Springer-Verlag.

Received January 20, 2011
Revision received June 27, 2011
Accepted June 27, 2011 ■