Journal of Cognitive Psychology

ISSN: 2044-5911 (Print) 2044-592X (Online) Journal homepage: http://www.tandfonline.com/loi/pecp21

Vocal mistuning reveals the origin of musical scales

Peter Q. Pfordresher & Steven Brown

To cite this article: Peter Q. Pfordresher & Steven Brown (2017) Vocal mistuning reveals the origin of musical scales, Journal of Cognitive Psychology, 29:1, 35-52, DOI: 10.1080/20445911.2015.1132024

To link to this article: http://dx.doi.org/10.1080/20445911.2015.1132024

JOURNAL OF COGNITIVE PSYCHOLOGY, 2017
VOL. 29, NO. 1, 35–52
http://dx.doi.org/10.1080/20445911.2015.1132024

Routledge
Taylor & Francis Group

Vocal mistuning reveals the origin of musical scales

Peter Q. Pfordresher[a] and Steven Brown[b]

[a]Department of Psychology, University at Buffalo, State University of New York, Buffalo, NY, USA; [b]Department of Psychology, Neuroscience and Behaviour, McMaster University, Hamilton, ON, Canada

ABSTRACT

Theories of the origin of tonality from the time of Pythagoras onward have assumed that the intervals used in musical scales are defined mathematically based on harmonic ratios. Virtually all such theories are predicated on tunable instruments (e.g. strings), whereas the voice is the most ancestral and universal instrument used to make music. In the present study, we analysed the tuning of sung musical intervals from a familiar song, doing so across both trained and untrained singers. Contrary to the predictions of traditional theories, we found that sung intervals (unlike those of instruments) showed marked overlap with neighbouring interval categories. Furthermore, we found that listeners of these sung productions did not base their aesthetic judgments of singing quality on the precision of tuning of sung intervals. We consolidate these results into a model of tonality based on both vocal and sensory factors that contribute to the formation of sung melodies.

ARTICLE HISTORY
Received 15 July 2015
Accepted 7 December 2015

KEYWORDS
Tonality; tuning systems; singing accuracy; scales; birdsong

Tonality refers to music's system of pitch relations, including the sequential arrangement of pitches that comprise musical scales, as well as the intervals that arise between both adjacent and non-adjacent pitches as a result of this arrangement. It also refers to the hierarchical arrangement of the pitches within a scale, such that certain scale-tones occur with greater frequency and have greater stability than others (Krumhansl, 1990; Large, 2010). Tonal systems vary across musical traditions cross-culturally, although scales often comprise five to seven pitches per octave. Western scales are predominantly diatonic, comprising only semitones and whole tones, although scales from many world regions contain minor thirds as well (e.g. the Arabic world, Turkey, India, East Asia).

A central question for both music theory and music psychology – which is the focus of the present research – is how the pitches of scales are derived. A theory of scales needs to account not just for Western music but for tonal systems throughout the world. There is a long tradition in musical practice dating back to the ancient Greeks for musical scales to be specified a priori and for instruments to be tuned based on these tonal principles. Such principles are based on the *mathematical ratios* of the fundamental frequencies of the tonic intervals (i.e. intervals relative to the tonic pitch) and/or melodic intervals (i.e. non-tonic intervals) of the scale (Loy, 2006). For example, the dominant tuning system in Western music since the seventeenth century, called equal temperament, is based on all 12 semitones within an octave having an identical frequency ratio, namely 1.0595:1 (a difference equalling 100 *cents*). This equivalence of semitones was adopted because it maintains a stable degree of tuning when a performer changes keys. Earlier tuning systems were not as flexible, but compensated for this limitation by emphasising the "purity" of musical intervals, which in mathematical terms means creating intervals comprised of small-integer ratios, such as 3:2 and 4:3. Such a system in which intervals are kept as mathematically pure as possible is called "just" intonation (for reviews see Burns, 1999; Handel, 1989; Thompson, 2013). Unfortunately, just intonation, unlike equal temperament, only allows for "pure" tuning for the key in which the instrument is tuned, and key changes can lead to a noticeable loss of intonation.

A musical preference for pure intervals is not simply an intellectual choice but a reflection of the

CONTACT Peter Q. Pfordresher pqp@buffalo.edu

acoustics of vibrating objects. Pitched sounds are made up not just of single frequencies but instead families of frequencies – called harmonics – that are related to one another as consecutive integer ratios of the fundamental frequency (2:1, 3:1, 4:1, etc.; Helmholtz, 1877/1954). In the natural harmonic series, one finds the musical intervals of just intonation, present as either adjacent harmonic ratios (e.g. 2:1 for the octave, 4:3 for the perfect 4th) or non-adjacent harmonic ratios (e.g. 5:3 for the major 6th, 15:8 for the major 7th). Theories of scale structure based on "harmonicity" (referred to here as *harmonicity theories*) argue that all of the ingredients of scales are found in single pitches (Gill & Purves, 2009; Parncutt, 1989; Terhardt, 1984). In other words, our sense of tonality is derived from the implicit harmonicity contained within pitched sounds.

However, the observation that certain musical intervals can assume the form of pure ratios found in the harmonic series does not mean that musical intervals *must only* assume these ratios. As we have noted, the standard system of tuning in our time represents a departure from pure ratios and yet music still sounds "in tune" to us. While tuning is often talked about abstractly in mathematical terms, there are important empirical questions about tuning that need to be addressed, including the extent to which musicians actually achieve tuning targets during performance as well as the extent to which perceivers of music are able to detect deviations from intended tunings. In addition, while all mathematical theories of tuning are strongly based on instruments that can be externally tuned, such as strings and pipes, the voice is the most universal musical instrument, and yet there is minimal empirical research regarding whether vocal tuning conforms to mathematical ratios the way that instrumental tuning presumably does, not least in indigenous cultures that lack academic treatises about idealised scales.

We propose that modern-day musical scales did not arise originally from a desire to maximise the purity of tuned intervals. Instead, we suggest that scales originated as a way of categorising pitches given intrinsic constraints in both the producers and perceivers of melodies. Tuning of intervals, though important now, may not have mattered as much in early forms of music (and in fact may not matter as much for modern music as one might expect, as we will show). Present-day tuning standards may instead reflect constraints that occurred when people started using tunable instruments and used these instruments to form harmonic intervals. At this point, sensitivity to sensory consonance may have led to standards of tuning that were not necessary when music was performed in unison, or by individuals. Our idea is not entirely novel. For instance, in their classic text, Dowling and Harwood (1986) observed that modern-day scales do not necessarily represent the earliest musical systems, observing that "… both equal temperament and small integer ratios arise from attempts to rationalize existing traditional scales" (p. 106). We here propose a basis for these traditional scales.

Our present concern is with the tuning of sung *melodic intervals* produced without accompaniment. A large literature exists concerning "intonation" (tuning) in musical performance that is dominated by studies of single-pitch matching, including both vocal and instrumental production (see Morrison & Fyk, 2002, for a review). We restrict our focus to sung intervals for the following reasons. First, as described above, the voice is the form of production that most likely possesses the features that constrained the formation of tonal structures to begin with. Second, we wish to look at the role of tuning independent of any peripheral constraints. The production of harmonic intervals or the presence of an accompaniment may influence interval production through beat tones associated with upper harmonics, based on peripheral sensory mechanisms (Plomp & Levelt, 1965). By contrast, melodic intervals produced without accompaniment reflect more purely the representation of intervals as ratios in schematic memory for tonal relationships (cf. Krumhansl, 1990; Krumhansl & Cuddy, 2010).

Thompson (2013) suggested that melodic intervals may reflect the joint constraints of harmonicity (as described above) and demands associated with grouping successive pitches, as in auditory scene analysis (Bregman, 1990). This proposal holds insofar as listeners respond to the consonance of successive pitches with the same sensitivity as they do simultaneous pitches. However, it is not clear that the perception of melodic intervals exhibits a fine-grained resolution. The discrimination of melodic intervals based on categorical perception of relative pitch exhibits discrimination thresholds on the order of 25–60 cents (see analyses reported by Burns & Ward, 1978; Smith, Kemler Nelson, Grohskopf, & Appleton, 1994), a dramatic difference from the fine-grained discrimination thresholds (on the order of 3 cents) found in simple pitch

discrimination (see Oxenham, 2013 for a review). Moreover, judgments of whether sung melodic intervals are "in tune" reveal thresholds on the order of 60 cents, thus crossing the boundary between adjacent interval categories (Hutchins, Roquet, & Peretz, 2012). It is therefore far from clear whether the perception of melodic intervals reflects the sensitivity to tuning found in harmonic intervals. We further suggest that vocal tuning during production may offer clues as to the expectations listeners have in perceiving melodic intervals.

As mentioned earlier, our research focus deals specifically with sung melodic intervals. Given that scales likely emerged from constraints associated with singing rather than instrumental performance, the tuning properties of the voice provide the most direct evidence about both the production and perception of melodic intervals. As in the aforementioned perception literature, existing research on vocal tuning suggests greater imprecision than would be predicted by harmonicity theories. However, these studies have been limited by their focus on expert performers and in some cases by the use of harmonic intervals. Vurma and Ross (2006) had professional singers vocalise ascending and descending minor seconds (m2), tritones, and perfect fifths in equal temperament. The average error across the three intervals was only 4 cents flat, with a standard deviation of 22 cents. Even for harmonic intervals, similar levels of imprecision have been found. Hagerman and Sundberg (1980) looked at barbershop singers, who are thought to employ just intonation in order to reduce beating between pitches. The interval error varied with the interval size and with the singer's role in the quartet, but the majority of interval errors were less than 20 cents. Similar levels of imprecision were reported by Devaney, Mandel, Ellis, and Fujinaga (2011) for professional performances of Schubert's *Ave Maria* (see also Devaney & Ellis, 2008).

In recent years, research on untrained singing has increased. In contrast to the previous studies of highly trained singers, Pfordresher and Brown (2007) looked at the imitative production of ascending and descending equal-tempered major thirds (M3), perfect fourths (P4), and perfect fifths (P5) in non-musician subjects. They found the mean interval errors to be 50–100 cents flat in the subjects categorised as "accurate" based on their criterion (see "Methods") and 100–250 cents flat in those subjects categorised as "poor-pitch singers" (as based on single-pitch-level imitative accuracy). A second study with non-musicians looked not just at pitch-matching tasks but at the singing of familiar songs from memory (e.g. "Happy Birthday") in order to examine internalised interval schemas in long-term memory (Pfordresher, Brown, Meier, Belyk, & Liotti, 2010). The mean interval error was 119 cents flat for the imitation samples and 73 cents flat for the familiar songs for intervals present in both stimulus sets.

A limitation of the studies summarised in the previous section is that their analyses conflate two levels at which produced intervals may miss the ideal and therefore be "out of tune". We adopt a measurement distinction, mentioned earlier, to address this shortcoming. The first level involves whether a sung interval is closer to some other target interval than the one that was intended (e.g. a perfect 5th [700 cents] is sung closer to a perfect 4th [500 cents]). Such category-based deviations are referred to as *semitone* deviations, and focus simply on intervals as discrete categories. The second level relates to the proximity of the produced interval to *any* appropriately tuned interval (e.g. a perfect 5th is sung as 680 cents instead of 700 cents, hence being 20 cents flat). Such fine-grained deviations are referred to as a *microtuning deviations* and address whether singers tune intervals based on an idealised template. In the case of trained singers, one might justifiably assume that sung intervals approximate the intended targets, but such an assumption may not hold in all cases.

Figure 1 illustrates the role of these two levels. The microtuning deviations in this figure represent the absolute difference between the sung interval (in cents) and the *closest possible* equal-tempered interval. By contrast, semitone deviations reflect the distance between the ideal interval that most

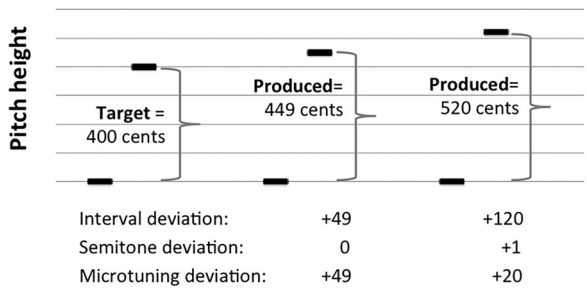

Figure 1. Illustration of a target interval and two incorrect productions of it that involve microtuning deviations (both examples) and a semitone deviation (rightmost example). Horizontal grid lines mark 100 cent increments.

closely matches the present produced interval and the intended target interval. Figure 1 shows a target interval of 400 cents (a major 3rd) followed by 2 attempted reproductions. In the first example, the produced interval would be considered accurate in terms of the semitone deviation metric, given that the closest ideal interval is the target interval. However, this interval is produced with a large microtuning deviation of 49 cents. In the second example, the produced interval is considered an error (the closest matching interval is 500 cents, a perfect 4th), but the microtuning deviation is smaller. Thus it is possible in principle for a participant to sing "incorrectly" (wrong intervals) and yet be more "in tune" vis-à-vis that interval.

Although nobody to our knowledge has compared microtuning with semitone deviations directly, this distinction does exist in the literature. Many studies of human singing categorise sung intervals discretely as erroneous or correct, which is similar to our semitone deviation measure (e.g. Dalla Bella, Giguère, & Peretz, 2007). For microtuning deviations, we draw on a study of birdsong by Araya-Salas (2012), although the author did not use this specific term. He classified intervals produced in the songs of nightingale wrens with respect to how close they were to the nearest acceptable melodic interval within just intonation. Differences were expressed as percentages, with 100% reflecting maximal distance, that is, halfway between two tonic intervals. Araya-Salas used this analysis to compare the tuning of melodic intervals in birdsong with the tuning of intervals produced by musical instruments having flexible pitch (e.g. violin, trombone).

The results of his analyses revealed striking differences between birdsong and instrumental music. Whereas instrumentally produced intervals had very low microtuning deviations, the distribution of microtuning deviations in birds exhibited only a negligible tendency toward accurate microtuning. Araya-Salas (2012) interpreted this mistuning of the birds' singing as indicating that, unlike humans, birds are not "musical". This was, of course, based on the assumption that the tuning of intervals is a defining property of music, which is consistent with the harmonicity view described earlier. By contrast, we propose that musical scales arose based on constraints in producers and listeners, and that the production constraints were specifically vocal, rather than instrumental. One can reasonably ask if Araya-Salas's comparison between instrumental performance in humans and vocal performance in birds is a fair one. It is an open question whether human singers are more similar to human instrumentalists or bird singers when it comes to microtuning. In addition, many professional musicians in Western culture go through an extensive process of academic training. Therefore, the human analogue of a songbird might not be a professionally trained singer but perhaps an individual without formal musical training who learns to sing as part of a communal process of socialisation. Like songbirds, human infants typically spend a great deal of time singing, and early vocal production reflects a blend of song-like and music-like features (Moog, 1976).

We were interested in putting Araya-Salas's method to the test in order to compare human singers with both bird singers and human instrumentalists in terms of their intonational precision when producing musical intervals. To do so, we took Araya-Salas's analytical method for examining microtuning and applied it to recordings of the singing of a familiar song from memory. We analysed both untrained and professionally trained singers, and divided untrained singers into those exhibiting accurate versus inaccurate pitch-matching abilities (cf. Pfordresher & Brown, 2007). The inclusion of these three groups was theoretically motivated. The comparison between trained singers and untrained accurate singers was designed to determine the extent to which explicit training in singing tuned intervals, as opposed to implicit learning based on passive exposure to music, influences production. The inclusion of inaccurate singers was designed to observe how a manifested inability to sing with consistent accuracy influences the microtuning of intervals, as opposed to full-on pitch errors (semitone deviations). Given Araya-Salas's sceptical assessment of bird musicality based on microtuning, we were interested in seeing if humans are significantly better than tuneful-sounding birds when it comes to the intonation of intervals when singing familiar melodies. If human singers look like Araya-Salas's instrumental musicians, then it supports the case for the existence of a strong species difference in musicality. If, on the other hand, human singers turn out to be as imprecise as birds with regard to microtuning, then this would have significant implications for theories of tonality and tuning in human music.

Methods

Recordings

Recordings from two groups of human singers were drawn from pre-existing corpora in which the participants sang "Happy Birthday" in a key of their choice. Despite the fact that "Happy Birthday" contains some complex features (most notably a non-tonic starting tone and an internal octave jump), it remains the best-recognised and most consistently sung melody among adult participants.

Untrained singers comprised 37 participants from Pfordresher and Brown (2007, Experiment 1). For the purposes of the present study, untrained singers were separated into two groups: *Accurate* singers and *VPID* singers, where VPID stands for "vocal pitch imitation deficit" (Pfordresher & Larrouy-Maestri, 2015), also referred to as "poor pitch singing" in previous publications (Pfordresher & Brown, 2007; Welch, 1979a, 1979b). Of these singers, 10 comprised all the VPID participants from Pfordresher and Brown (2007, Experiment 1). The remaining participants were the first 27 accurate participants from that study.[1] We did not include all accurate singers from that study here ($N = 69$) as that would have created an extreme imbalance across sample sizes. The mean age of the participants was 23 years, 21 participants (57%) were male, and the mean years of training on a musical instrument was 0.8.

The recorded corpus for the *trained singers* was retrieved from an online resource (reported in Larrouy-Maestri & Morsomme, 2014b). Participants in this sample were classically trained opera singers from the French-speaking region of Belgium (Liège). Each participant performed "Happy Birthday" in French in both an expressive (operatic) style and an unexpressive ("flat") style. Although we analysed both performance styles, results here focus on the unexpressive style of singing, which better resembles the performance style of the untrained singers. We measured recordings of the first 16 singers from a total of 50 singers in the corpus. The mean age of the performers in the group was higher than that of the untrained singers ($M = 36.94$), due in part to their years of vocal training. Of the trained singers sampled, all but two were female.

Pitch analysis

Initial pitch estimation

The continuous fundamental frequency (F0) signal was extracted from each recording using the autocorrelation method in Praat (Boersma & Weenink, 2013). Segmentation of each syllable was performed by hand as annotations in Praat. Although the performances varied in language (English and French) and in the personal names used in "Happy Birthday" (English performances were directed to the experimenter: either Julie, Erik, or Danny; French performances were directed to Pauline), segmentation was performed so that the ordering of syllables was matched across performances.

Following F0 extraction and segmentation, a MATLAB script was used to compute a single pitch value for each syllable. For each sample, the median F0 value was computed from the middle 50% of a syllable. We used this middle portion in order to minimise the influence of scoops at the beginning and end of a syllable as well as possible artefacts in F0 extraction (e.g. use of median rather than mean). All performances that were included contained the correct number of syllables.

Computation of interval deviations and error scores

Interval-error scores were computed in the following way. Each successive pair of performed notes was transformed into a difference in cents via a logarithmic transformation of the frequency ratio. The resulting vector of pitch intervals (in cents) was compared to the corresponding vector of intended intervals based on equal temperament. We computed *interval deviation* scores using the procedure described by Pfordresher et al. (2010) based on the difference in the absolute value of the produced intervals from ideal performance. This calculation leads to negative values for intervals sung smaller than intended and to positive values for intervals sung larger than intended, and gives the same results for ascending and descending intervals. Interval deviations that exceeded ±50 cents (i.e. a 1-semitone window) were considered errors.

[1] Pfordresher and Brown (2007) categorized as VPID any singer for whom the absolute value of mean pitch deviations exceeded 100 cents. Since that time, this criterion has been criticized as being too lax with respect to how many participants may be categorized as accurate, with most researchers now advocating for a 50-cent criterion (see Pfordresher & Larrouy-Maestri, 2015 for discussion). We chose to follow the categorization used in the original paper for purposes of comparison. Nevertheless it is worth noting that 23 of our 27 accurate participants (85%) had error pitch deviation scores between 0 and 50 cents and could thus be considered accurate under either criterion.

We used equal temperament rather than just intonation (used by Araya-Salas, 2012) as our reference system for intervallic tuning due to the dominance of equal temperament in the music our singers were exposed to during their lifetime. Thus, to the extent that singers implicitly tune to a fixed interval template, equal temperament is the most likely template they would use. Practically speaking, the differences between the equal temperament and just tuning systems are subtle enough that the results reported here would not differ significantly if they were reported using just intonation.

Decomposition of interval deviations into semitone and microtuning deviations

As described in the Introduction, the deviation scores computed as described above are influenced both by microtuning and the accuracy of intervals in a discrete sense. The *microtuning* score for each sung interval was the absolute difference between the sung interval (in cents) and the *closest possible* equal-tempered interval. These values could thus range from 0 (most in-tune) to 50 cents (most mistuned) and could vary continuously between these values. In contrast to microtuning deviations, *semitone deviations* were computed using the absolute difference between the nearest acceptable interval and the *target* interval. These scores were computed in semitones (integer values) rather than cents. Each semitone deviation is thus a discrete value, although means across intervals for a given participant varied on a continuous scale.

Distributions of microtuning deviations

The technique described in Araya-Salas (2012) was based on computing a "percent proximity" measure for microtuning deviations and then plotting distributions of this measure. We computed similar measures of microtuning for human singing as a means of creating a comparison with his data. Microtuning deviations are expressed as inverse percentages of the farthest possible distance from perfect tuning. For the cents scale, the largest deviation from tuning is an absolute tuning error of 50 cents (sharp or flat). In the Araya-Salas measure, such a deviation would be expressed as 0% proximity. By contrast, perfect tuning (a deviation of 0 cents) would be expressed as 100% proximity. We adapted this procedure using Equation (1):

$$\mathrm{IP}_x = \left[1 - \frac{min\{(|x| mod 100), 100 - (|x| mod 100)\}}{50}\right] \times 100 \quad (1)$$

where IP is the interval proximity for sung interval x, expressed as a percentage. First we transform the absolute value of the interval using modulo 100 division, which yields a deviation from pure tuning in equal temperament (spaced in 100 cents) that ranges from 0 to 99. The min function in Equation (1) causes that deviation to range from 1 to 50, reflecting its proximity to a lower or higher purely tuned interval. Finally, the deviation is expressed as an inverse proportion of the total deviation possible (50 cents) to reflect its percent proximity to the closest possible tonic interval relative to the maximum possible microtuning deviation of 50 cents, with higher values indicating closer proximity. Following Araya-Salas, we generated frequency distributions of these proximity scores, using five equally spaced bins ranging from 0% to 100%. We compared distributions based on human data to distributions based on the published bird data of Araya-Salas (2012) using graphical estimates from Figure 4 of that paper, which contained data from the three birds that showed the greatest tendency toward microtuning and three instrumental musical performances.

Ratings of singers

In addition to analyses of singing data, we also report subjective evaluations of singing accuracy for the untrained singers from an unpublished study conducted at the University of Texas at San Antonio in 2005. Raters were 24 students taking Introduction to Psychology. There were 14 females and 10 males, ranging in age from 18 to 34 years old ($M = 20.5$). Most participants had some type of musical experience ($M = 4.5$ years of experience), although no participant was an expert in singing or singing pedagogy.

On each trial, raters heard one singer's recording of "Happy Birthday". Afterwards, they were told to rate its pitch quality on a Likert scale, with 7 meaning "accurate" and 1 meaning "inaccurate". Raters were instructed to focus specifically on the accuracy of the melody's pitch and to ignore vocal timbre and timing. Means across raters were then

used to determine which acoustic variables best predicted the subjective evaluations.

Results

Microtuning of sung intervals resembles birdsong, not instrumental music

We first report results for the microtuning of sung intervals, in other words the proximity of sung intervals to the nearest possible equal-tempered interval. As a starting point, we incorporated the analysis procedure described in Araya-Salas (2012; see "Methods"). For ease of comparison, we restricted the data on trained singers to those trials in which participants were instructed to sing without expression, as this led to productions that were most comparable to the style used by the untrained singers (e.g. minimal use of vibrato).

Figure 2 presents a comparison between human and songbird production, where the histograms represent the original data for bird singing and human instrumental performance from Araya-Salas' (2012) publication and the three lines represent the three groups of human singers. We analysed the frequency distributions from Figure 2 using chi-square analyses. We adjusted the critical alpha level to 0.01 due to the necessity of multiple comparisons (each corpus group with two different distributions, and three pairwise comparisons between corpus groups = 3 shared comparisons for each group). We first used chi-square goodness-of-fit tests within each of the three human datasets in order to determine whether the exhibited patterns deviated form uniformity (i.e. a flat distribution), where uniformity implies minimal conformity with equal-tempered tuning. The distribution for the trained singers deviated significantly from uniformity, $\chi^2(5) = 46.52$, $p < .01$, suggesting a tendency toward microtuning in equal temperament. However, neither group of untrained singers differed from uniformity [accurate $\chi^2(5) = 2.96$, VPID $\chi^2(5) = 0.74$, critical value = 13.28 for $p < .01$]. We next ran chi-square tests of independence to compare the frequency distributions among the three classes of singers. The distribution of microtuning for trained singers differed from both accurate untrained singers, $\chi^2(5) = 20.10$, and VPID untrained singers, $\chi^2(5) = 32.51$. However, not surprisingly, the two untrained singer groups did not differ from one another, $\chi^2(5) = 0.17$.

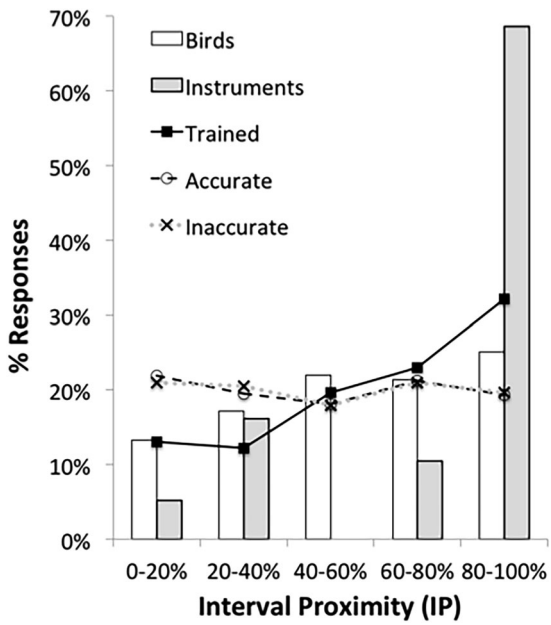

Figure 2. Distribution of microtuning for melodic intervals using the interval proximity metric (Equation (1)). Three groups of singers performing "Happy Birthday" are plotted along with the original songbird data (white bars) and instrumental data (grey bars) from Araya-Salas (2012). Data from the earlier paper are reproduced with permission of the author. Note that, because the two groups of untrained singers performed similarly, their data points overlap.

While these results demonstrated that professional training in singing is associated with an improvement in microtuning compared to no formal training, microtuning in every human group better resembled the songbird data than the instrumental data from Araya-Salas (2012), which argues against Araya-Salas's conclusion that birdsong differs from human musicality due to inaccurate microtuning. None of the three human groups produced microtuning distributions that differed from songbirds [trained singers, $\chi^2(5) = 3.00$, accurate untrained, $\chi^2(5) = 3.08$, VPID untrained, $\chi^2(5) = 4.23$]. By contrast, all three groups deviated from the frequency distribution exhibited by instrumental performers [trained singers, $\chi^2(5) = 56.54$, accurate untrained, $\chi^2(5) = 103.18$, VPID untrained, 82.46], with instrumental performances showing a markedly stronger tendency toward equal temperament. Thus, while extensive professional training in singing leads to a stronger tendency toward "in-tune" singing than the absence of such training, it does not increase this tendency beyond what has been observed in putatively non-musical birds, despite the clear intention of human singers to perform within a rigorously specified tuning system.

Microtuning is a weak predictor of singing ability

As described earlier, each interval deviation was decomposed into two components: a microtuning deviation and a semitone deviation. The boxplots shown in Figure 3 represent the distribution of these scores across all individuals in each group, with each data point being the average for a single participant. Note that, although the semitone deviation for a given interval can only be an integer value, the mean of these values for a participant usually falls in between integers. Statistical analyses compared untrained accurate singers, untrained VPID singers, and trained singers, the latter using the unexpressive ("flat") style (a single between-subjects factor). However, for comparison purposes, the boxplots also show the data from trained singers while singing expressively (i.e. the operatic style).

The similarity between the two untrained groups in the previous analysis (Figure 2) stands in stark contrast to the large differences that these groups exhibit in interval deviation scores (e.g. Pfordresher & Brown, 2007). Thus, it seems reasonable to expect that semitone deviations, rather than microtuning deviations, underlie such group differences. This expectation was validated in the analysis of semitone deviations shown in Figure 3(a). The main effect of group was significant with a large effect size, $F(2, 50) = 31.05$, $p < .001$, $\eta_p^2 = 0.55$. More importantly, Tukey's Honestly Significant Difference (HSD) tests verified that all pairwise differences were significant, including the difference between accurate untrained and VPID untrained singers. Thus, VPID singing is characterised by errors in selecting the appropriate interval class, whereas mistuning may be a feature that is characteristic of all singing that has not been guided by formal training in the Western classical tradition.

In contrast to this, analyses of microtuning (Figure 3(b)) showed far smaller group differences. The effect of group (VPID singers, accurate untrained singers, trained singers not using expression) was significant, albeit with an effect size roughly half of that for semitone deviations, $F(2, 50) = 10.16$,

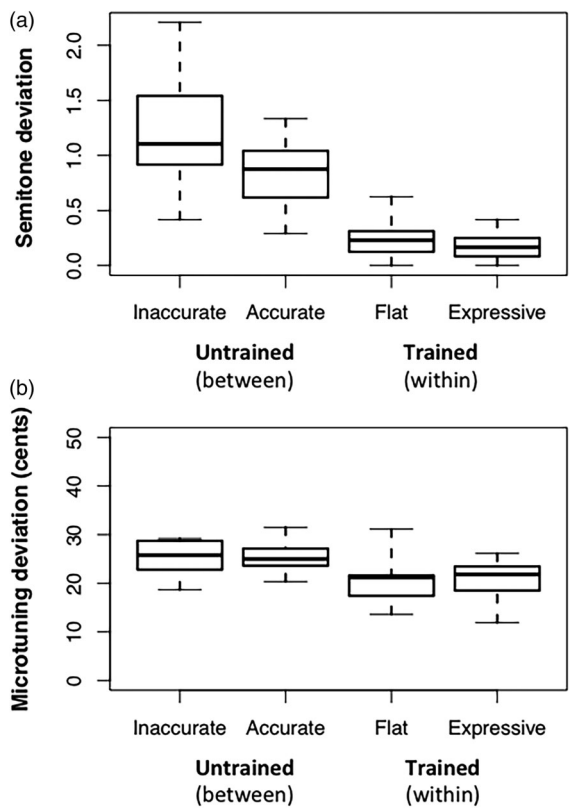

Figure 3. Boxplots representing distributions of semitone deviations (a) and microtuning deviations (b). Averages were computed across all syllables of "Happy Birthday" for an individual. Rectangles show the inter-quartile range, with the internal horizontal line representing the median. Whiskers represent the total range of data.

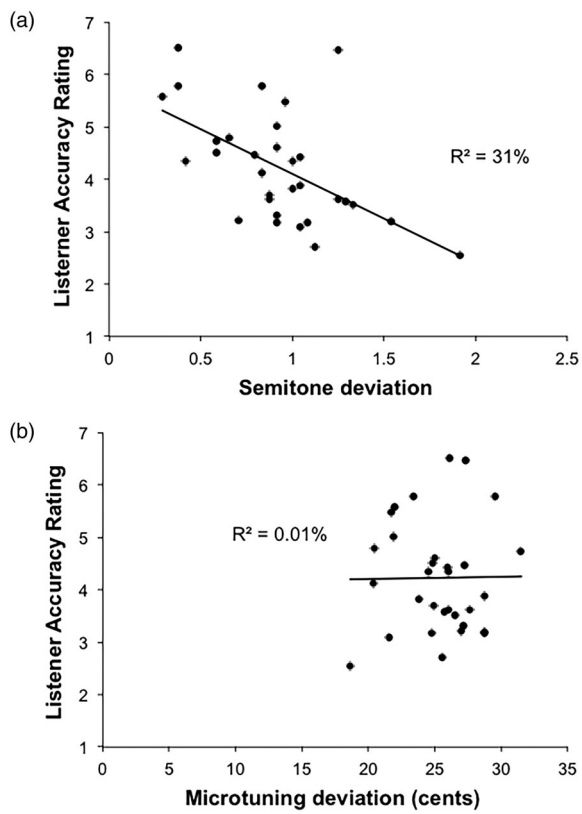

Figure 4. The relationship between the mean rating of pitch accuracy by listeners (ordinate for both panels) with the mean semitone deviation (a) or mean microtuning deviation (b) generated by singers.

$p < .01$, $\eta_p^2 = 0.29$. Post-hoc comparisons (Tukey's HSD, $\alpha = 0.05$) revealed that microtuning deviations were lower for trained singers (not using expression) than either untrained singer group. However, in contrast to semitone deviations, the difference between untrained groups was not significant. Interestingly, trained singers exhibited a somewhat higher degree of mistuning while singing with expression than without it, consistent with research concerning the effect of expressivity on singing accuracy (Larrouy-Maesteri & Morsomme, 2014a, 2014b).

One limitation of this analysis is that group categorisation for the untrained singers was based on analyses of acoustic production data for trials involving the imitation of short, unfamiliar melodies. Not surprisingly, performance on these imitation tasks is correlated with accuracy in the reproduction of familiar songs (Berkowska & Dalla Bella, 2013; Pfordresher & Brown, 2007; Pfordresher et al., 2010; Wise & Sloboda, 2008). As such, the predictor variable (group) was not entirely independent of the outcome variable (microtuning or semitone deviation). For this reason, we used the production data from the untrained singers (only) to perform a correlational analysis to measure how well microtuning and semitone deviations predict listener ratings. As can be seen in Figure 4(a), the relationship between accuracy ratings from listeners and mean semitone deviations for singers was strong and significant, $r(29) = 0.56$, $p < .001$. By contrast, listener ratings had a negligible relationship with microtuning deviations, $r(29) = 0.10$, $p > .20$ (Figure 4(b)). Thus, listener evaluations of singing quality appear to be based in large part – perhaps even exclusively – on semitone deviations, rather than the degree of microtuning.

Interval categories are broad and overlapping

Next we consider how the apparent failure to vocally tune intervals influences the distinctiveness of different interval-classes in sung performance. Given that singers show a relatively weak tendency toward accurate microtuning (cf. Figure 2), one may rightly wonder how it is that tonal information may be communicated at all in song.

In order to investigate this, we examined the accuracy of the most frequent intervals of "Happy Birthday". Figure 5 shows frequency distributions for particular interval-classes independent of their melodic context, that is, where any given interval may be preceded or followed by several alternatives. As can be seen, these distributions are very wide, overlapping quite strongly with neighbouring interval-classes. This is seen most strikingly for untrained singers (Figures 5(b) and 5(c)), where the overlap is so strong that it is difficult to see where category boundaries should even exist. Note that the offset of the mode for the unison interval is due to the fact that singers occasionally produced the word "Happy" as a semitone instead of a

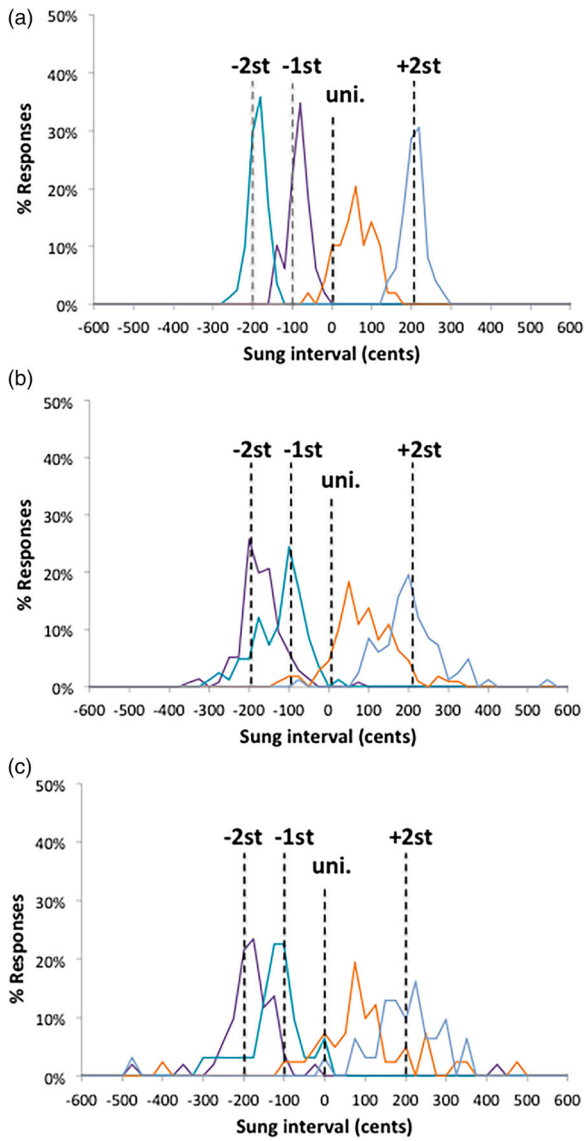

Figure 5. Frequency distributions of sung intervals along with dashed lines representing ideal tuning within the equal-tempered scale for trained singers (a), accurate untrained singers (b), and VPID untrained singers (c). Intervals plotted are the most frequently occurring within "Happy Birthday". Abbreviations: st, semitone; uni, unison.

unison, where the first syllable acted as a leading tone.

Singers preserve ordinal differences in interval size

Because of limitations in the precision of vocal-motor control intervals taken outside of a melodic context may be variable enough to blur category boundaries, as clearly seen in Figure 5. The question remains whether intervals in a *sequential* context reflect a tendency to avoid overlap between successive intervals. In other words, singers might adjust the size of their intervals on a pitch-by-pitch basis to preserve contrasts in interval size in a sequential context. In order to look further into these differences between isolated and sequential intervals, we developed a measure to quantify the maintenance of sequential categorical distinctions in singing, referred to here as *ordinal integrity*. The logic of this measure is that it compares the ordinal change in interval size across two successive melodic intervals – either "increasing", "decreasing" or "uniform" changes in interval size between successive intervals – independent of direction (contour). This constitutes an ordinal code for changes in interval size (i.e. an interval contrast) that can take on the values of + 1, −1, or 0, where "uniform" interval pairs are operationally defined as a change equal to or smaller than 50 cents. This ordinal code is then compared between a given performance and the target melody. Ordinal integrity is operationalised as the number of interval contrasts that match between performance and target.

Figure 6 presents an illustrative example using musical notation (in practice, of course, pitch intervals often fall in between notated intervals, as we have shown). In the target melody, shown on the left, absolute interval sizes (in semitones, ST) are compared with respect to ordinal change in size. Thus, from the first interval (C–E, size = 4 ST) to the second interval (E–A, size = 5 ST), the ordinal change is an increase in size (+). By contrast, for the transition from the second interval to the third interval (A–C, size = 3 ST), the ordinal change is a decrease in size (–) compared to the previous one (5 ST). The resulting ordinal code creates a representation of whether the change across successive intervals is an increase or decrease in interval size, disregarding both the absolute magnitude and direction of these changes. The "poor" reproduction of this melody (shown on the right side of Figure 6) maintains the melodic contour perfectly but fails to maintain the ordinal integrity of the interval sizes. For example, the second interval (F–G, size = 2 ST) is smaller than the first interval (C–F, size = 5 ST), in contrast to the ordinal code for the target.

We measured ordinal integrity in this way for each performance in our corpora, and compared this measure to the percent of individual intervals that were sung accurately (based on interval-error counts, described in the "Methods"). Because this accuracy measure does not take into account relationships across successive intervals, it measures interval accuracy in a context-free sense. Figure 7 plots "percent correct" measures for both ordinal integrity and interval accuracy. The ANOVA yielded a main effect of group, $F(2, 50) = 42.08$, $p < .001$, $\eta_p^2 = 0.63$, and measure, $F(1, 50) = 56.81$, $p < .001$, $\eta_p^2 = 0.48$, with the latter indicating generally better performance on the ordinal integrity measure than interval accuracy. The group × measure interaction was not significant ($p = .13$, $\eta_p^2 = 0.08$). However, planned pairwise comparisons across measures were significant for trained singers, $t(15) = 3.39$, $p < .01$, and untrained accurate singers, $t(26) = 6.36$,

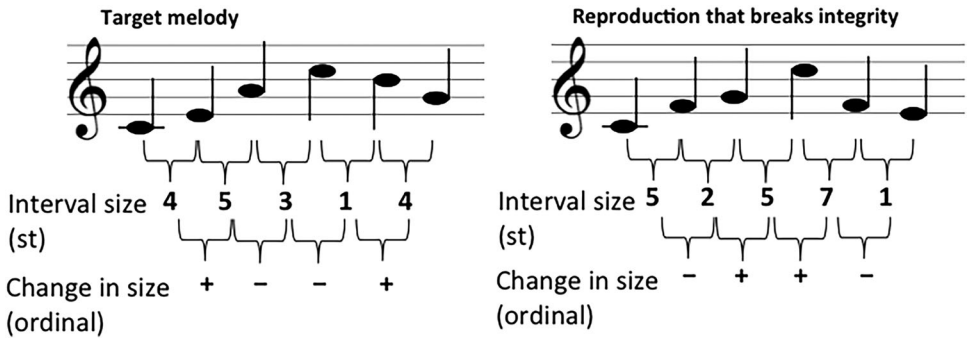

Figure 6. An example of the ordinal code of a target melody (left) and a reproduction that fails to retain ordinal integrity of interval size, even while contour is accurate (right).

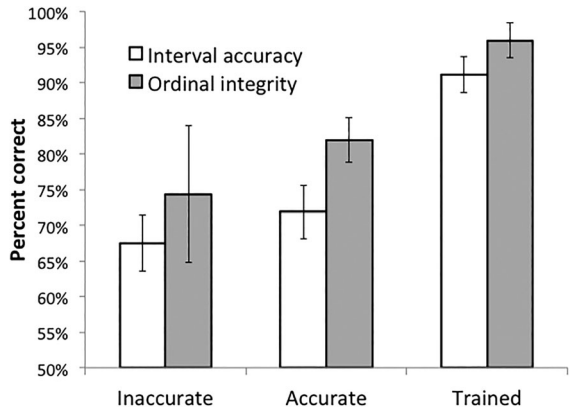

Figure 7. Percent correct for ordinal integrity and interval accuracy measures. Error bars represent 95% confidence intervals.

$p < .001$, but not VPID singers ($p = .10$). Thus, singers in general may uphold category boundaries within a sequential context, possibly at the expense of the integrity of interval categories when examined outside of a melodic context. Melodic intervals may thus be represented ordinally rather than on an interval scale.

Discussion

Interval tuning is imprecise in both production and perception

The corpus analyses that we reported here verified the hypotheses we derived from earlier research. Contrary to the predictions of harmonicity theories and ideas based on low-level frequency-processing properties of the auditory system, singers do not reliably tune intervals to an accepted standard (here, from equal-tempered tuning). Instead, singers produce intervals that broadly overlap adjacent interval-classes, to the point that the populations of produced intervals should be hard to distinguish. Training had some influence on the accuracy of this microtuning, but to a very small extent compared to accuracy at the semitone level. And even here, the advantage of training was qualified by singing style. Even more dramatically, microtuning in human singing was strikingly similar to tuning observed in the nightingale wren by Araya-Salas (2012), who used his results as the basis for the claim that the singing of nightingale wrens failed to meet the criterion of musicality seen in human instrumentalists. Furthermore, listeners seem to discount mistuning in their evaluations of singing accuracy.

What can be made of these results? One possible interpretation is that the results found here reflect nonlinearities in the auditory system (cf. Large, 2010). However, such remarkable mistuning may surpass the kind of generosity to mistuning predicted by such a system. Moreover, we believe that a good starting point for understanding the origins of tonal systems lies in production. Communication, through both music and language, comes about from interactions between producers and perceivers. The job of the perceptual system is to decode the intentions of producers. It thus makes sense to treat as foundational those limitations that are inherent in production, to which the perceptual system must adapt in order to decode producers' intentions.

Based on the imprecision of produced melodic intervals observed in this study, we propose that the internal representations of musical intervals are better conceptualised as "islands" of frequency ratios than as the singular points (i.e. specific harmonic ratios). For such an analysis, we can take our lead from the study of phonetics, where vowels are shown to exist as extended regions in a vowel space defined by the first two formants of the speech signal. Rather than being points in this space, or lines relating a constant relationship between F1 and F2, the formant relationships for perceived vowels are shaped like long ellipses in this two-dimensional vowel space (Hillenbrand, Getty, Clark, & Wheeler, 1995; Klein, Plomp, & Pols, 1970; Peterson & Barney, 1952; Turner & Patterson, 2003). In other words, a vowel is not a single F1-to-F2 frequency ratio but instead an island of frequencies extending hundreds of Hz along each dimension. Just as speakers create repertoires of vowels so as to achieve contrastive distinction among them, so too do singers attempt to create contrasts among pitches when creating intervals during melody formation. Hence, spacing principles are critical, and singers have to navigate between islands in order to achieve pitch contrasts in creating intervals. In support of this, we showed that distinctions among intervals were blurred when seen out of context (Figure 5), whereas listeners showed a marked tendency to maintain ordinal contrasts among neighbouring intervals (Figure 7). Hence, what singers ultimately try to achieve is a spacing of pitches during melody formation, and this can be achieved by intervals taking on a wide range of values within a single melody, as dependent upon vocal factors and melodic context.

Studies on the categorical perception of melodic intervals, mentioned in the Introduction, offer further support for this conceptualisation. Although categorical perception of melodic intervals has been reported, the perceptual resolution of intervals is much poorer than would be expected based on harmonicity theories, again suggesting the idea that interval categories are broad in perception. Moreover, only musically trained listeners have exhibited categorical perception for isolated intervals (Burns & Ward, 1978). Non-musicians require the presence of melodic context, either a real or imagined context (e.g. Smith et al., 1994). The fact that interval categories are coarse-grained and – for non-musicians – dependent on context coheres with our view that scales are based on a system that is oriented toward the limitations of producers. Whereas the resolution of the auditory system is highly fine-grained, the ability to recognise melodic-interval categories is more closely aligned with the imprecision seen in interval production.

Toward a vocal theory of tonality

One stream of thought in the field of music theory since the time of Pythagoras has started from the point of mathematical theories of tuning and then *imposed* these theories onto musical practice under the assumption of naturalness and purity. Importantly, these observations (starting with Pythagoras) were inspired not by the singing voice but by physical properties of tunable instruments (strings in the case of Pythagoras). However, the present-day music cognition literature suggests a persistent uneasiness concerning the link between harmonicity and tonality. Even Helmholtz (1877/1954), who is commonly associated with harmonicity, argued that, "… the system of Scales … does not rely solely upon unalterable natural laws but is also, at least partly, the result of esthetical [*sic*] principles …" (p. 235). Dominant theoretical accounts of tonal hierarchies reflect this separation. They typically emphasise the importance of category structures, and therefore bypass the role of tuning (e.g. Balzano, 1980; Krumhansl, 1990; Krumhansl & Cuddy, 2010). We propose that the origins of tonality may lie in the *mistuning* of melodic intervals.

As with phonetics, the analysis of musical scales should be based on the observed patterns of vocal melody production, and this information should be used to *infer* scales *a posteriori* based on the manners in which the voice moves intervallically in pitch space. Scales are essentially abstractions of the way people create melodies, as based on patterns of interval spacing. As Arom, Fernando, and Marandola (2007) note with regard to the *a capella* choral singing of the Bedzan Pygmies of central Africa:

> For native musicians the musical scale does not exist in and of itself, i.e. independently of the repertoire to which it applies. It only exists in its execution, and a study of the scales can only be based on *pieces* drawn from the traditional repertoire. (p. 107, italics in the original)

This is not to say that Pygmy singers do not have tuning, only that this tuning does not in any sense precede performance, much as in the case of a singer in Western culture who lacks a training in music theory.

Figure 8 shows that scales can be conceptualised in two related ways. Historically, they have been defined as a graded series of intervals expressed with respect to their relationship to a tonic, that is, as tonic intervals (Figure 8(b)). This kind of representation is standard in music theory; for example, the major scale can be represented as M2, M3, P4, P5, M6, M7, P8. The other way is a sequential arrangement of adjacent melodic intervals that can be combined to form any tonic interval (Figure 8(a)); for example, the major scale can be represented as the sequence M2, M2, m2, M2, M2, M2, m2. Obviously, these two representations are interrelated and interconvertible. The tonic-interval model is particularly well-suited for instruments, since they can be tuned a priori to a tonic standard. However, this is not the case with the voice. The voice can sing a wide range of pitches, but there is no sense in which the voice can be tuned to a scale in preparation for singing. In addition, in traditional cultures that do not have academic tuning theories or a strong presence of tunable instruments, scales develop through melody production and though vocal imitation of other singers. Hence, a theory of scales based on tonic intervals is not a reasonable model for a non-tunable instrument like the voice. In fact, the model we present next bears striking resemblance to measurements of vocal tuning among the Bedzan Pygmies, whose polyphonic choral music is primarily sung *a capella* and who have no explicit terms for scales in their language. Arom et al. (2007), in line with our results with Western singers, demonstrated that the tuning of the pentatonic scale in Pygmy

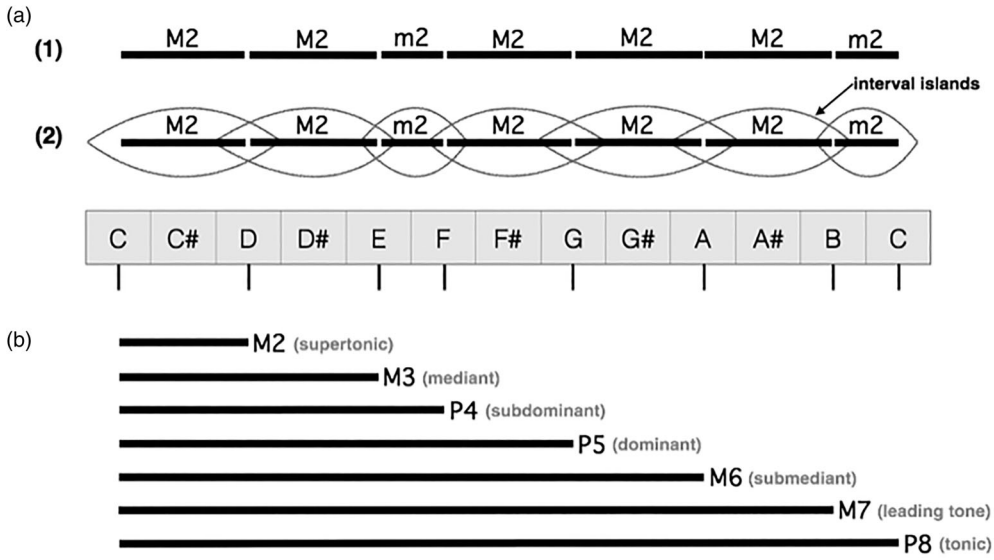

Figure 8. Contrasting conceptions of a scale as either (a) a sequential arrangement of melodic intervals, or (b) a graded arrangement of tonic intervals. The 12 pitches of the chromatic scale are shown at centre and apply to both Panels A and B. The subset of seven pitches making up the diatonic major scale are marked with vertical tick marks below the respective pitches. Line (1) of Panel A shows the arrangement of melodic intervals of the diatonic major scale, and Line (2) shows our concept of "interval islands" superimposed upon this as oval-shaped structures. This forms the basis of our "interval-spacing" model of scales. Panel B shows the series of seven tonic intervals that make up the diatonic major scale. Abbreviations: m2, minor second; M2, major second; M3, major third; P4, perfect fourth; P5, perfect fifth; M6, major sixth; M7, major seventh; P8, octave.

singers is extremely imprecise, showing very wide interval categories completely comparable with our interval islands. We strongly believe that ethnographic analyses of singers in traditional cultures should form the basis of a theory of scales, not the mathematical treatises of the West.

Line (2) of Figure 8(a) shows the melodic-interval model with schematics of interval islands superimposed upon the melodic intervals, as based on the empirical data presented here and elsewhere about the broad tuning of musical intervals in both production and perception. We call this the *interval-spacing model of scales*, and argue that scales are based on achieving an optimal level of spacing among pitches such that these pitches are distinguishable from neighbouring pitches in both production and perception. Such distinguishability depends critically upon melodic context, including the contour and size of the preceding and proceeding intervals. We hypothesise that tone languages should operate according to the same interval-spacing principle as described here, since tones have to be constantly distinguishable from neighbouring tones across a diversity of melodic contexts in a sentence, for example as the voice declines in pitch towards the end of a sentence (Ladd, 1984). So, tone languages should be another general example of interval-spacing principles in action, although this occurs in the *absence* of scales and tonality (as may have been the case in the earliest music). Tone languages, despite their name, are not tonal in the musical sense (Patel, 2008). This suggests that interval-spacing principles should operate across music and speech, and that tonality is simply one manifestation of this principle in which additional mechanisms are involved, such as recurrent pitch classes in melodic ascent and descent.

The interval-spacing model considers both vocal-motor and perceptual factors in explaining the nature of scales (schematised in Figure 9), in contrast to the traditional reliance on perceptual factors alone (e.g. harmonicity and consonance). The inadequacy of sensory factors alone in explaining scales is seen in the observation that the discrimination threshold for pitch is extremely small, on the order of 3–5 cents when using pure tones. Based on such a fine-grained threshold, musical scales could theoretically contain as many as 400 divisions per octave. Yet, in reality, scales contain two orders of magnitude fewer pitches that that, with the vast majority of scales throughout the world being based on 7 or fewer pitches. Hence, other factors

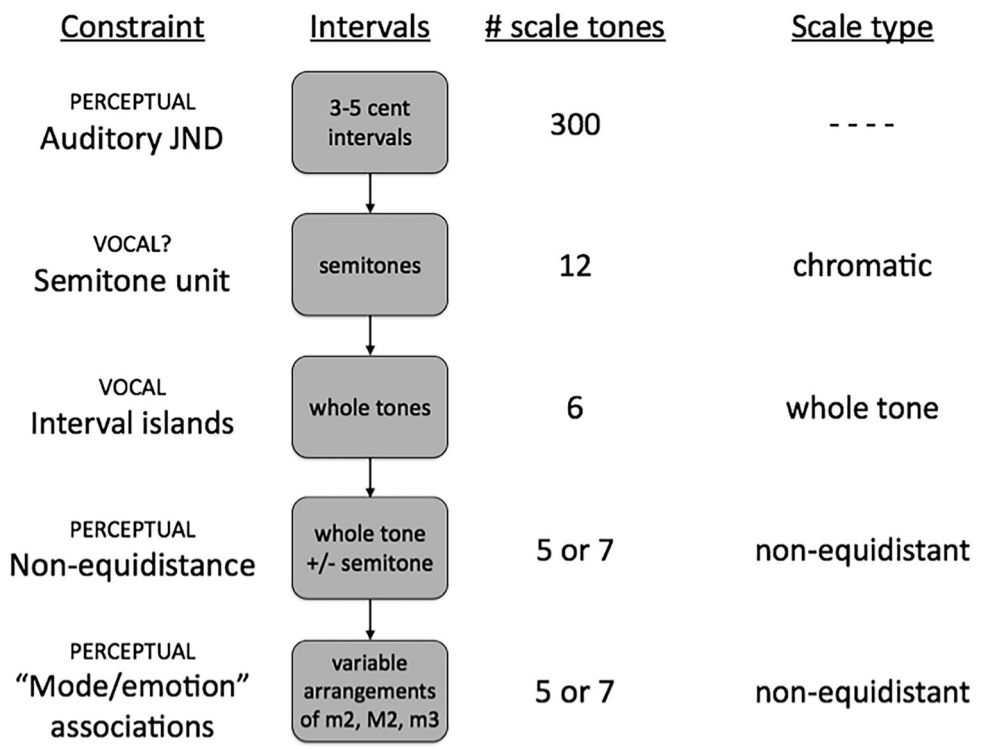

Figure 9. Conceptual development of an interval-spacing model of tonality. See text for details. Abbreviations: JND, just noticeable difference; m2, minor second; M2, major second; m3, minor third.

must be at play in constraining the number of scale pitches to less than a dozen per octave.

Burns (1999) and others have theorised about how the equal-tempered chromatic scale might represent the most reliable division of the octave into equal units. While Burns does not talk about the physiological basis of this semitone constraint in music, we propose that it might represent the smallest reliable unit of phonatory movement at the level of the laryngeal muscles, hence the smallest unit that is reliably singable. Whatever its basis, Burns' hypothesis brings the lower limit of musically usable intervals from the 3–5 cents of pitch perception to the 100 cents of equal-tempered semitones, thereby effecting a significant reduction in the number of scale pitches from potentially 400-note scales to 12-note scales. However, 12-note scales are extraordinarily rare outside of the Western chromatic scale, and so there must be other factors operating to reduce the potential of using the semitone as the major unit for scale creation and melody generation.

We propose that one critical factor is *the imprecision involved in vocally generating intervals*, as shown in our demonstration of interval islands. An important outcome of this observation for a theory of scales is that the whole tone, rather than the semitone, is the more distinguishable interval for singing, and that the whole tone scale is a more vocally reliable scale than the chromatic scale. This emphasis on whole tones might provide a reasonable explanation for the universal human inclination to create scales containing 6 ± 1 pitches (Justus & Hutsler, 2005; Stevens & Byron, 2009).

Unfortunately, the stability of the intervallic spacing of the whole tone scale for singing is offset by its equidistant nature. Scales are overwhelmingly non-equidistant, comprised of more than one step size per octave, generally two or three. Such non-equidistance permits the existence of a tonic pitch and an associated tonal hierarchy that is impossible to achieve with an equidistant scale (cf. Balzano, 1980). So, instead of the whole tone scale, we see a generative principle by which scales are built up of combinations of melodic intervals that are "whole tones +/− a semitone", namely semitones, whole tones, and minor thirds. Scales comprise sequential combinations of these three step-sizes, and they vary based on the sequences by which these three intervals are recombined. Finally, a given culture typically contains not one but several scale types, each one differing in the sequential arrangement of its melodic intervals. This diversity in scale types may be driven by variable "mode/emotion" associations, in other words

differing emotional connotations – and thus communicative meanings – of each scale (Huron, 2006, 2008; Parncutt, 2014; Temperley & Tan, 2013).

In proposing a specific representation of intervals here, we are not suggesting that all melodies are encoded in interval-specific ways. Although a familiar melody like "Happy Birthday" (used here) is likely to be encoded in terms of its specific intervals, it has long been known that more unfamiliar melodies are encoded with respect to their scale (key) and melodic contour (Dowling, 1978; Dowling & Bartlette, 1981). It is important to note that a scale-based memory representation, according to our view, would simply apply the same level of imprecision to the representation of pitch classes within the scale. Moreover, our primary concern is with the likely origin of scales, rather than the way in which listeners encode pitch information in memory.

Vocal tuning in song and speech

Gill and Purves (2009) argued that the evolution of tonality in music occurred due to an exposure to the harmonic sounds of speech. In focusing on the spectral aspects of sounds, Gill and Purves are aware that patterns of change in the fundamental frequency of the voice during speech do not conform to the tuning principles of scales. This leads to a curious paradox in their argument. Why would music, *but not speech itself*, take advantage of the harmonicity of speech-sounds to generate tonality? Even languages that linguists refer to as "tonal" are not at all tonal in the musical sense. We would like to turn this argument on its head and make a different point about the evolutionary connection between music and speech. Unlike a violin or flute, the voice is used not just for music but also for speech. This probably has the effect of making the singing voice much more speech-like than is the case for a musical instrument. In other words, the voice is not a dedicated musical instrument since it double-duties for speech. Even though people strive for level tones and perfect intervals when singing, their voices actually perform as an imprecise musical instrument. Singing words and having to deal with a stream of changing syllables may make the voice into a much less precise instrument from the standpoint of microtuning than violins and flutes that do not have to concern themselves with changing their resonator properties to create different timbres on each note. So, our suggestion of a "speech mode" of vocal music is that either text itself or the need to repeatedly change vocal tract configurations or the combination of the two serves to compromise the microtuning properties of the singing voice, as compared to instruments. While music unquestionably evolved as a vocal phenomenon, vocal music might have the properties that it does because of a physiological connection of the voice with speech. If this is so, then mathematical theories based on instruments might provide an unrealistic representation of tonality, and what Araya-Salas found for songbirds might be exactly what we would expect for the human voice.

In addition, speech and song cannot help but be intertwined. All singing requires some vowel or another for production; much singing is word-based (rather than vocable-based); and the patterning of fundamental frequency in speech – while clearly *not* tonal in the musical sense – is highly melodic, rather than being monotonic or unpitched. People do not speak in monotones but instead navigate through pitch space according to rather standard melodic formulas (Cruttenden, 1997; Ladd, 1996). Beyond that, both speech and song are acquired through vocal imitation. People do not tune their voices when they want to sing, but instead either match occurrent melodies if music is being played or produce melodies from memory based on previous episodes of imitative learning. Finally, there are a host of vocal and stylistic factors that work to compromise intervallic tuning during singing, including vibrato, portamento, melisma, and many stylistic devices that are used by singers to alter the timbre of the voice, through nasality, glottal shake, creaky voice, noise, wobble, and the like. Moreover, vocal timbre varies across vocal registers (Sundberg, 2013). Many of these stylistic devices strongly increase the emotional and aesthetic appeal of sung music (Friberg, Bresin, & Sundberg, 2006). However, thanks to the fact that listeners accept more imprecision in sung timbres than instrumental timbres ("vocal generosity", Hutchins et al., 2012), they can often do so without creating the impression of mistuning, as we found in our own rating study (Figure 4(b)).

Limitations and future directions

There were several limitations of the present study. First, the analysis of the instrumental data was from Araya-Salas's (2012) analysis of recorded music and not from our own studies. It will be important to

conduct laboratory studies of instrumental tuning using a large sample of musicians, much as was done with our singing analysis. Next, our vocal study relied on a single song, "Happy Birthday". Moreover, as was shown in the interval analysis (Figure 7), there were certain idiosyncrasies associated with that song, particularly with respect to the unison. So, it will be important to analyse additional songs to corroborate the results with "Happy Birthday". Finally, whereas we propose the foundation of a new framework in which to conceptualise musical scales, we have not proposed a systematic model that generates scales, as has been proposed by harmonicity theorists (e.g. Gill & Purves, 2009). As such, the present work can be considered the first step in a new direction, as an alternative to the already well-developed (though possibly misdirected) harmonicity view. We would simply add in this regard that one tradition in music theory that shows compatibility with our viewpoint is the set-theoretic approach (Hanson, 1960). By specifying constraints on vocal spacing – for example, favouring adjacent whole tones in a scale, or avoiding adjacent semitones or minor thirds – it should be possible to use combinatoric principles to elaborate possible scales that satisfy the spacing constraints on adjacency. The resultant scales may not have the highest harmonicity, but they might be the ones that are most reliably singable.

An important application of this work is to musical development. Children are well-known to sing "out of tune", and tuneful singing develops only by about age 11, if at all (Welch, 2006). Because tuning is specified a priori in Western culture, children's singing is defined as out of tune with reference to it. However, it will be important to apply the analysis of microtuning developed in this paper to children's singing, and to investigate the relative frequency of semitone deviations and microtuning deviations in their productions. Given that the dominant developmental model of singing is that children make a transition from an out-of-tune precursor to a crystallized in-tune style of singing (Welch, 2006), then it will be important to characterise the nature of the tonal system that is employed at this precursor stage (and presumably into adulthood for some people). Since young children seem to have a restricted vocal range compared to older children and adults (Welch, 1979b, 2006), interval-spacing principles might provide critical insight into how the tonal properties of singing develop over the course of childhood. In addition, given our contention that singing is characterised by a "speech mode" of production, it will be important to compare microtuning when singing is done with and without words (cf. Berkowska & Dalla Bella, 2009; Mantell & Pfordresher, 2013; Racette & Peretz, 2007; Welch, 1979a).

Finally, an important application of this work on tonality is to examine vocal scales from a cross-cultural perspective and to apply the same type of analysis to music that has been applied to phonetics, namely inferring scales from patterns of production (Arom et al., 2007), instead of specifying them a priori on theoretical grounds. Virtually all cross-cultural research on scales has focused on instrumental tunings (Daniélou, 1999; Ellis, 1885), and has been carried out in the tradition of mathematical theories of frequency ratios. The only way to generate a universal theory of scales is to develop a cross-cultural research programme into the tuning properties of sung scales, especially in indigenous cultures that have few melodic instruments and that do not have written treatises on music theory. As Lomax (1968) pointed out, vocal production-styles for singing vary dramatically across cultures, and this could have a significant impact on microtuning, most especially for noisy singing styles. Anecdotally, we ourselves have heard numerous recordings of indigenous singers that have sounded intervallically imprecise and even out of tune to our Western ears. Historically, music from such cultures has been labelled as "primitive" and non-musical by previous generations of scholars, on exactly the same grounds that Araya-Salas deemed wren song to be non-musical. Therefore, a big point of contention in looking at sung scales in indigenous cultures will be in knowing what the *intended* pitches are supposed to be, just as with birdsong. We can talk about the mistuning of singing relative to "natural" intervals, but these sung intervals might actually represent the intended pitches of the singer, human or animal. It would be a big mistake to equate tonality with one particular system of tuning, no matter how natural it might seem. As we have argued in this paper, the true test of naturalness is cross-cultural universality, not mathematical abstraction.

Conclusions

We analysed the tuning of sung musical intervals in both highly trained and untrained singers, and found that the intervals were extremely broad, showing marked overlap with neighbouring interval

categories. This provides a challenge for theories of tonality based on specific ratios and perfect intervals. We argue instead that a theory of tonality should be based on the sensorimotor properties of vocal production and melody formation. According to this view, intervals are more like islands than single ratios, and melody generation is about navigating between these islands to maintain the distinguishability of pitches over the course of a melody (or sentence). Such a view has important applications to childhood singing development, speech production for tone languages, and cross-cultural analyses of tuning systems, most especially for vocal music. It also has strong evolutionary implications for the origins of both singing and speaking.

Acknowledgements

We thank Patrick Savage, Michael Hall, Steven Morrison, David Temperley, Carol Krumhansl, Steven Demorest, Timothy Hubbard and three anonymous reviewers for helpful comments on a previous version of the manuscript. We also thank Steven McAdams for pointing out the link between our work and that of Arom and colleagues.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by National Science Foundation grant BCS-1256964 to PQP and from the Natural Sciences and Engineering Research Council (NSERC) of Canada to SB.

References

Araya-Salas, M. (2012). Is birdsong music? Evaluating harmonic intervals in songs of a Neotropical songbird. *Animal Behaviour*, *84*, 309–313.

Arom, S., Fernando, N., & Marandola, F. (2007). An innovative method for the study of African musical scales: Cognitive and technical aspects. In C. Spyridis, A. Georgaki, G. Kouroupetroglou, & C. Anagnostopoulou (Eds.), *Proceedings of the 4th sound and music computing conference* (pp. 107–116). Athens: University of Athens.

Balzano, G. J. (1980). The group-theoretic description of 12-fold and microtonal pitch systems. *Computer Music Journal*, *4*, 66–84.

Berkowska, M., & Dalla Bella, S. (2009). Reducing linguistic information enhances singing proficiency in occasional singers. *Annals of the New York Academy of Sciences*, *1169*, 108–111.

Berkowska, M., & Dalla Bella, S. (2013). Uncovering phenotypes of poor-pitch singing: The sung performance battery (SPB). *Frontiers in Psychology*, *4*, 714.

Boersma, P., & Weenink, D. (2013). Praat: Doing phonetics by computer (Version 5.4.09) [Computer program]. Retrieved August 25, 2014, from http://www.praat.org/

Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound*. Cambridge, MA: MIT Press.

Burns, E. M. (1999). Intervals, scales, and tuning. In D. Deutsch (Ed.), *The psychology of music* (2nd ed., pp. 215–264). San Diego, CA: Academic Press.

Burns, E. M., & Ward, W. D. (1978). Categorical perception - phenomenon or epiphenomenon: Evidence from experiments in the perception of melodic musical intervals. *Journal of the Acoustical Society of America*, *63*, 456–468.

Cruttenden, A. (1997). *Intonation* (2nd ed.). Cambridge: Cambridge University Press.

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.

Daniélou, A. (1999). *Introduction to the study of musical scales*. New Delhi: Munshiram Manoharlal.

Devaney, J., & Ellis, D. P. W. (2008). An empirical approach to studying intonation tendencies in polyphonic vocal performances. *Journal of Interdisciplinary Music Studies*, *2*, 141–56.

Devaney, J., Mandel, M. I., Ellis, D. P. W., & Fujinaga, I. (2011). Automatically extracting performance data from recordings of trained singers. *Psychomusicology*, *21*, 108–136.

Dowling, W. J. (1978). Scale and contour: Two components of a theory of memory for melodies. *Psychological Review*, *85*, 341–354.

Dowling, W. J., & Bartlett, J. C. (1981). The importance of interval information in long-term memory for melodies. *Psychomusicology*, *1*, 30–49.

Dowling, W. J., & Harwood, D. L. (1986). *Music cognition*. San Diego, CA: Academic Press.

Ellis, A. J. (1885). On the musical scales of various nations. *Journal of the Society of Arts*, *33*, 485–527. Reprinted in K. Kaufman Shelemay (Ed.), *Garland Library of Readings in Ethnomusicology 7* (pp. 1–43). New York: Garland, 1990.

Friberg, A., Bresin, R., & Sundberg, J. (2006). Overview of the KTH rule system for musical performance. *Advances in Cognitive Psychology*, *2–3*, 145–161.

Gill, K. Z., & Purves, D. (2009). A biological rationale for musical scales. *PLoS One*, *4*, e8144.

Hagerman, B., & Sundberg, J. (1980). Fundamental frequency adjustment in barbershop singing. *Journal of Research in Singing*, *4*, 3–17.

Handel, S. (1989). *Listening: An introduction to the psychology of auditory events*. Cambridge, MA: Bradford Books/MIT Press.

Hanson, H. (1960). *Harmonic materials of modern music: Resources of the tempered scale*. New York: Appleton-Century-Crofts.

von Helmholtz, H. (1877/1954). *On the sensations of tone as a physiological basis for the theory of music*. New York: Dover.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, *97*, 3099–3111.

Huron, D. (2006). *Sweet anticipation: Music and the psychology of expectation*. Cambridge, MA: MIT Press.

Huron, D. (2008). A comparison of average pitch height and interval size in major-and minor-key themes: Evidence consistent with affect-related pitch prosody. *Empirical Musicology Review*, *3*, 59–63.

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

Justus, T., & Hutsler, J. J. (2005). Fundamental issues in evolutionary psychology of music: Assessing innateness and domain specificity. *Music Perception*, *23*, 1–27.

Klein, W., Plomp, R., & Pols, L. C. W. (1970). Vowel spectra, vowel spaces and vowel identification. *Journal of the Acoustical Society of America*, *48*, 999–1009.

Krumhansl, C. (1990). *Cognitive foundations of musical pitch*. Oxford: Oxford University Press.

Krumhansl, C. L., & Cuddy, L. (2010). A theory of tonal hierarchies in music. In M. R. Jones, R. R. Fay, & A. N. Popper (Eds.), *Music perception* (pp. 51–88). New York: Springer.

Ladd, D. R. (1984). Declination: A review and some hypotheses. *Phonology Yearbook*, *1*, 53–74.

Ladd, D. R. (1996). *Intonational phonology*. Cambridge: Cambridge University Press.

Large, E. W. (2010). A dynamical systems approach to musical tonality. In R. Huys & V. K. Jirsa (Eds.), *Nonlinear dynamics in human behavior* (pp. 193–211). Berlin: Springer-Verlag.

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

Larrouy-Maestri, P., & Morsomme, D. (2014b). Effects of melody and technique on acoustical and musical features of western operatic singing voices. *Journal of Voice*, *28*, 332–340.

Lomax, A. (1968). *Folk song style and culture*. Washington, DC: American Association for the Advancement of Science.

Loy, G. (2006). *Musimathics: The mathematical foundations of music*. Volume 1. Cambridge, MA: MIT Press.

Mantell, J. T., & Pfordresher, P. Q. (2013). Vocal imitation of speech and song. *Cognition*, *127*, 177–202.

Moog, H. (1976). *The musical experience of the pre-school child* (C. Clarke, Trans.). London: Schott.

Morrison, S. J. and Fyk, J. (2002). Intonation. In R. Parncutt & G. McPherson (Eds.), *The science & psychology of music performance: Creative strategies for teaching and learning* (pp. 183–198). Oxford: Oxford University Press.

Oxenham, A. J. (2013). The perception of musical tones. In D. Deutsch (Ed.), *Psychology of music* (3rd ed., pp. 1–33). Amsterdam: Academic Press.

Parncutt, R. (1989). *Harmony: A psychoacoustical approach*. Berlin: Springer-Verlag.

Parncutt, R. (2014). The emotional connotations of major versus minor tonality: One or more origins? *Musicae Scientiae*, *18*, 324–353.

Patel, A. D. (2008). *Music, language, and the brain*. New York: Oxford University Press.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, *24*, 175–184.

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

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.

Pfordresher, P. Q., & Larrouy-Maestri, P. (2015). On drawing a line through the spectrogram: How do we understand deficits of vocal pitch imitation? *Frontiers in Human Neuroscience*, *9*, 271.

Plomp, R., & Levelt, W. J. M. (1965). Tonal consonance and critical bandwidth. *Journal of the Acoustical Society of America*, *38*, 548–560.

Racette, A., & Peretz, I. (2007). Learning lyrics: To sing or not to sing? *Memory & Cognition*, *35*, 242–253.

Smith, J. D., Kemler Nelson, D. G., Grohskopf, L. A., & Appleton, T. (1994). What child is this? What interval was that? Familiar tunes and music perception in novice listeners. *Cognition*, *52*, 23–54.

Stevens, C., & Byron, T. (2009). Universals in music processing. In S. Hallam, I. Cross, & M. Thaut (Eds.), *Oxford handbook of music psychology* (pp. 14–23). Oxford: Oxford University Press.

Sundberg, J. (2013). The perception of singing. In D. Deutsch (Ed.), *The psychology of music* (3rd ed., pp. 69–106). London: Academic Press.

Temperley, D., & Tan, D. (2013). Emotional connotations of diatonic modes. *Music Perception*, *30*, 237–257.

Terhardt, E. (1984). The concept of musical consonance: A link between music and psychoacoustics. *Music Perception*, *1*, 276–295.

Thompson, W. F. (2013). Intervals and scales. In D. Deutsch (Ed.), *Psychology of Music* (3rd ed., pp. 107–140). Amsterdam: Academic Press.

Turner, , R. E., & Patterson, R. D. (2003). An analysis of the size information in classical formant data: Peterson and Barney (1952) revisited. *Journal of the Acoustical Society of Japan*, *33*, 585–589.

Vurma, A., & Ross, J. (2006). Production and perception of musical intervals. *Music Perception*, *23*, 331–344.

Welch, G. (2006). Singing and vocal development. In G. McPherson (Ed.), *The child as musician: A handbook of musical development* (pp. 311–329). Oxford: Oxford University Press.

Welch, G. F. (1979a). Poor pitch singing: A review of the literature. *Psychology of Music*, *7*, 50–58.

Welch, G. F. (1979b). Vocal range and poor pitch singing. *Psychology of Music*, *7*, 13–31.

Wise, K., & Sloboda, J. A. (2008). Establishing an empirical profile of self-defined 'tone deafness': Perception, singing performance, and self-assessment. *Musicae Scientiae*, *12*, 3–23.