JASA EXPRESS LETTERS ARTICLE asa.scitation.org/journal/jel

A reversal of the song advantage in vocal pitch imitation

Peter Q. Pfordresher[a)]
Department of Psychology, University at Buffalo State University of New York, Buffalo, New York 14260, USA
pqp@buffalo.edu

Abstract: Previous experiments have documented an advantage for vocal pitch-matching when participants sing back a short melody, in contrast to when participants attempt to imitate the pitch contour of spoken English. These results appear to confirm recent claims that music involves greater precision of pitch than speech. A re-analysis of these data is reported here that focuses on imitation of pitch trajectories within sung notes or spoken syllables. When analyzed this way, the domain-based difference reverses and speech imitation exhibits an advantage relative to song imitation. These results suggest that domain-specific advantages in imitation vary as a function of timescale.

[Editor: Charles C. Church] https://doi.org/10.1121/10.0009729

Received: 16 December 2021 **Accepted:** 14 February 2022 **Published Online:** 9 March 2022

1. Introduction

A strong prevailing view in music cognition is that musical pitch is processed with finer grained precision than spoken pitch (Patel, 2011, 2014). This view is articulated most commonly with respect to perception. Several lines of evidence inform this view. First, fundamental frequency (f_0), the acoustic correlate of pitch, is generally more stable across the duration of a musical note than across a spoken syllable [e.g., Zatorre and Baum (2012)]. Syllables are generally considered to have a roughly analogous role to notes given that, like musical notes, syllables convey rhythm, and musical text-setting typically maps notes onto syllables. This difference in stability is quite prominent when comparing speech to instrumental music, but also present when comparing speech to song, as I will show later. The greater stability of pitch (qua f_0) suggests that through exposure listeners may extract more reliable pitch information in musical contexts than in speech. Second, the rate of musical notes is generally much slower than the rate at which spoken syllables proceed (Ding *et al.*, 2017; Patel, 2014), which also benefits the extraction of pitch in listeners. Finally, listeners exhibit greater perceptual sensitivity to alterations of pitch and spectral fine structure for music than for speech (Albouy *et al.*, 2020; Sares *et al.*, 2018; Zatorre and Baum, 2012).

Such observations have led to the hypothesis that lateralization effects found in brain activity during listening to speech and music may reflect a trade-off between temporal precision (optimized in the left hemisphere) and spectral precision (optimized in the right hemisphere) (Albouy *et al.*, 2020; Zatorre *et al.*, 2002; Zatorre and Gandour, 2008). The idea that pitch precision is greater in perception of music than speech has also led to the hypothesis that musical training may benefit auditory perception of pitch in general (Kraus and Chandrasekaran, 2010; Moreno and Bidelman, 2014; Patel, 2011, 2014).

Several studies likewise suggest that the people imitate sung pitch more accurately than spoken pitch. Mantell and Pfordresher (2013) had participants listen to and then vocally imitate short melodies and sentences that were matched with respect to verbal content, global pitch contour and duration, but varied in other respects. Analyses of f_0 were based on vectors of all sampled values, and thus sensitive to tracking of time-varying properties of produced pitch. Sample-by-sample comparisons revealed greater mean absolute deviations between target and imitated f_0 values (i.e., poorer pitch matching) for speech than song that is independent of verbal content (Mantell and Pfordresher, 2013), amount of exposure (Wisniewski *et al.*, 2013), and explicit instructions to imitate pitch (Pfordresher *et al.*, 2021). A similar advantage was also found when analyses focused on mean f_0 within each sung note or spoken syllable, including the accuracy of matching single pitches (absolute pitch), and pitch intervals [relative pitch, Liu *et al.* (2013) and Wang *et al.* (2021)].

An important qualification of the evidence for a song advantage in production discussed above is that they all involve the proximity of produced to target f_0, rather than the similarity of the pattern of change in f_0 over time. The hypothetical example shown in Fig. 1(A) illustrates the limitation of such an approach. In this example, the reproduction of a 4-note melody is produced consistently sharp (too high). The average absolute difference between target and imitated

[a)]Author to whom correspondence should be addressed.

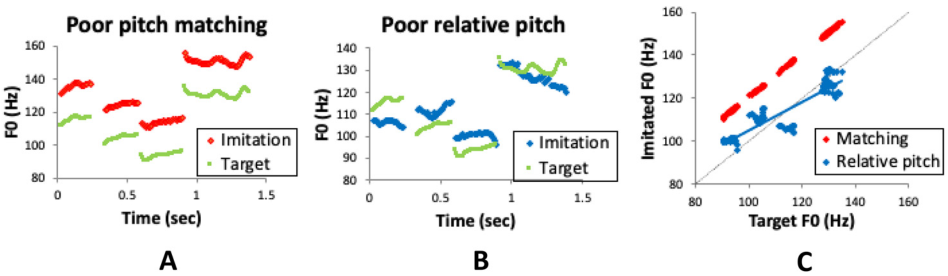

Fig. 1. Illustrative hypothetical examples of different forms of inaccurate singing. Panels (A) and (B) plot f_0 by time separately for hypothetical imitation and target patterns, whereas panel (C) plots the state space formed by correlating time-matched f_0 values for target (X) and imitation (Y). Colors in panel (C) match the corresponding color used to plot imitation values in panels (A) and (B), and the light black line highlights unity.

f_0 values across the sequence is 291 cents—almost a minor third in musical terms. However, it seems inappropriate to characterize such a reproduction as simply inaccurate given that the important relative pitch patterning—the changes in pitch from note to note and even from sample to sample—is preserved. By contrast, the hypothetical example in Fig. 1(B) illustrates a performance that is on average more proximal to the target (M absolute deviation = 101 cents), yet the performance would likely be perceived as less faithful to the target than the example in Fig. 1(A) due to relative pitch. For instance, the participant's mean f_0 increases slightly from the first to second note, in contrast to the target. Within individual notes, the patterning of produced f_0 also deviates from the target. For instance, there is a gradual decrease in produced f_0 across the duration of the final note that is not present in the target.

Mantell and Pfordresher (2013) introduced a measure that is sensitive to the vocal imitation of pitch patterning across the entire pitch trajectory, based on correlating produced to target f_0 across all samples, after adjusting the timing so that each sample matches a corresponding point in relative time. Examples illustrating this procedure are shown in Fig. 1(C). The points relating target to produced f_0 for Fig. 1(A), shown in red, reflect the perfect association with respect to pitch patterning for this performance ($r = 1.00$). By contrast, the relationship for the data in Fig. 1(B), shown in blue, yields a weaker association ($r = 0.89$).

The pitch correlation approach arose from a desire to capture imitation of relative pitch patterns on a large timescale, based on transitions across subsequent notes or syllables, as well as on a smaller timescale based on fluctuations of f_0 within notes and syllables. Small timescale fluctuations in pitch patterning are significant in speech, and appear to be perceptually salient in song as well (Larrouy-Maestri and Pfordresher, 2018). Analyses of production using pitch correlations for f_0 variability across an entire sequence have yielded no evidence for a song advantage in previous studies (Mantell and Pfordresher, 2013; Pfordresher *et al.*, 2021; Wisniewski *et al.*, 2013). By contrast, analyses of song versus speech imitation that average f_0 across the entire note or syllable tend to preserve the song advantage when relative pitch is analyzed based on pitch intervals (Liu *et al.*, 2013). Thus, the reduction of the song advantage appears to be linked to the role of imitating fluctuations in f_0 at a small timescale.

Here, I present a re-analysis of previous data using a technique that focuses on the imitation of pitch patterning specific to small-timescale fluctuations. Specifically, pitch correlations were computed within each event (a produced note or syllable), and the resulting set of correlations were averaged across all events in a sequence. The resulting measures reflect how well participants' production tracks localized changes in f_0 over time, independent of whether the producer accurately imitates transitions across event boundaries with respect to relative pitch.

2. Method

2.1 Materials

A sample of imitative productions from two previous studies were re-analyzed in the current study (Mantell and Pfordresher, 2013; Pfordresher *et al.*, 2021). Imitations were produced by 75 participants (n female = 33, M age = 19.49, SD age = 1.76), who participated in exchange for course credit. Participants had little formal musical training (M years of private instruction on instrument or voice = 2.67, SD = 4.10), and all were fluent speakers of English. Target stimuli that were imitated by participants comprised 12 short phrases of 3–5 syllables with matching text across speech and song. Stimuli were created by having a male and female model talker produce phrases as statements and then as questions, to vary the pitch contour in ways common to American English speech. Song stimuli were then created by composing short tonal melodies based on the global pitch contour of sentences, and having the same model talkers sing those melodies using the same text as for speech. Speech and song stimuli were thus matched with respect to phonetic properties and global pitch contour, while varying with respect to pitch stability, tonal encoding, rhythm and tempo [for further details, see Mantell and Pfordresher (2013)]. In addition, the items analyzed here were segmented into syllables based on syllable initial phonemes.

2.2 Procedure

On each trial, participants listened to a target stimulus without vocalizing, and then repeated the sequence as soon as possible after the target finished. For all trials analyzed here, participants repeated sequences with the intention of imitating its pitch [non-intentional imitation trials were dropped from Pfordresher *et al.* (2021)] and phonetic content ["wordless" trials were dropped from Mantell and Pfordresher (2013)], for each of the 48 items (12 phrases × 2 intonation contours × 2 domains). Some conditions from the previous data sets were dropped from the present analysis so that all participants would be analyzed based on a comparable task.

2.3 Data analysis

Analyses focused on the difference in the domain represented by the target, song or speech, aggregating across all other factors related to stimulus structure. I report four different analyses, two of which replicate the procedures used in past studies (though results are new, given the aggregation across data sets), and two new analyses based on segmenting into separate events. For the purposes of these analyses an event is considered a syllable, with syllables constituting divisions of subsequent notes for song stimuli. For all recordings (target and imitation) f_0 samples were extracted using PRAAT (Boersma and Weenink, 2013) and any octave errors were adjusted by altering the range of acceptable Hz values in PRAAT based on visual and auditory inspection of the extracted f_0 pattern [for further details, see Mantell and Pfordresher (2013)].

All analyses involve temporal matching of target to imitative performance on a sample-by-sample basis that adjusts for any differences in absolute time (total duration) and relative time (duration of events within the total duration), using custom MATLAB routines. Specifically, the duration of each syllable in the target was adjusted to match that of the imitation, and f_0 values for the target were re-sampled so that each sample matched a corresponding sample in the imitation based on its position within the total event duration. Note that in this process, all the resampling was performed on the target rather than imitation recording in order to preserve all information from the imitations. All sampled f_0 values were converted to cents (100 cents = 1 musical semitone) relative to the tonic tone for song stimuli (G2 for male participants, G3 for female participants).

Two sets of analyses varied with respect to focus on pitch *proximity* versus pitch *patterning* (analogous to absolute versus relative pitch), and reflect the distinction illustrated in Fig. 1. Pitch proximity analyses use the mean of absolute differences across every matched f_0 sample. This kind of analysis reflects pitch matching as shown in Figs. 1(A) and 1(B). Pitch patterning analyses use the Pearson correlation between matched target and imitation f_0 values, as shown in Fig. 1(C) and primarily reflect how changes in relative pitch in the target match corresponding changes in the imitation.

Analyses also varied with respect to timescale. Analyses based on a *within-event* timescale proceed by first extracting measures for each individual syllable in a sequence, and then averaging across those observations. The resulting measure (not used in previous papers) does not reflect any inaccuracies related to the transition across events in an imitation. Given the distinction between proximity and patterning described above, the use of a within-event timescale should only exhibit an effect on pitch patterning. Analyses based on a *cross-event* timescale involved first concatenating samples across all events to form a reconstituted sequence, and then computing relevant measures across every sample in the sequence. Note that previous studies only used analyses based on the cross-event timescale.

Figure 2 illustrates how different results can come out of pitch patterning analyses when using different timescales. The long panel in the top left displays a song target and an imitation across time [similar to Figs. 1(A) and 1(B)]. The top right panel shows the results of a pitch patterning analysis that uses the cross-event timescale [similar to Fig. 1(C) and previous studies]. As can be seen the association across target and imitated f_0 values is strong, leading to a significant positive correlation ($r = 0.97$). However, the correlations within each event, shown in the bottom row, are inconsistent and less positive; the average across these values (pitch patterning for the within-event timescale) is near zero ($r = 0.08$). This difference reflects the fact that the imitator does not consistently follow fine-grained changes in pitch within notes, but does accurately replicate the large pitch changes across notes. each trial, participants listened to a target stimulus without vocalizing, and then repeated the sequence as soon as possible after the target finished.

3. Results

Table 1 reports descriptive statistics for measures of imitative performance as a function of domain (song versus speech), timescale of analysis (cross-event or within-event), and focus of analysis (pitch proximity versus pitch patterning). The last three rows report statistics based on properties of imitated target stimuli which will be discussed later. Because the central issue of this report is on the advantage of song over speech for different measures and timescales of analysis, song advantage difference scores were constructed in the following way. For each participant and item that shared the same text setting and global contour (e.g., "He ate it all?" spoken as a question or sung as a melody with a rising pitch contour), differences were computed for each measure such that positive values indicate more accurate imitation for song (speech minus song for measures of pitch proximity, song minus speech for measures of pitch patterning).

As can be seen in Table 1, song advantage scores were all positive (reflecting a song advantage) with one exception: When pitch patterning accuracy is measured on a within-event timescale (highlighted in bold). Figure 3 plots

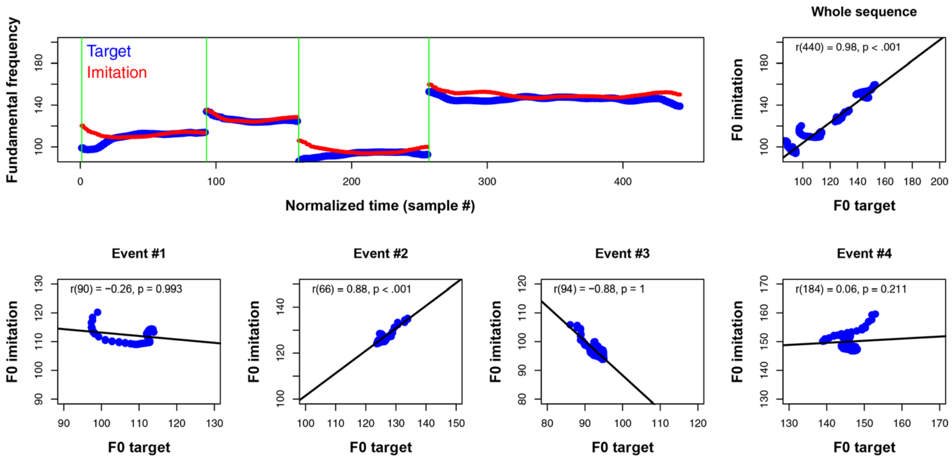

Fig. 2. Example of song imitation and pitch pattern accuracy measures based on cross-event and within-event timescales. Vertical lines in the top left panel reflect segmentation based on phonetic properties of the target and imitation, after aligning all onsets and event durations. Each panel in the lower row plots the association between target and imitated f_0 for each successive event.

differences in song advantage scores as a function of the focus of the analysis [Fig. 3(A) versus 3(B)], and timescale (abscissa grouping). Song advantage scores in Fig. 3 were averaged across all trials for an individual participant.

Differences in Song Advantage Scores as a function of Event Timescale were analyzed separately for each measure (pitch proximity, pitch patterning), using Linear Mixed Effects analysis with R (R Core Team, 2021) and lme4 (Bates *et al.*, 2015). Participant and item (i.e., the text setting and gender that served as targets across speech and song condition) were entered as random intercepts in a baseline model, and statistical significance was determined using Log Likelihood measures for models with Event Timescale as a fixed effect relative to this baseline (Baayen *et al.*, 2008). It is also important to establish whether song advantage scores within each Event Timescale differ from zero; this was established by analyzing the significance of intercepts for separate baseline models within each timescale.

The introduction of Event Timescale improved model fit to the pitch patterning measure [Fig. 3(A)] relative to the baseline model $\chi^2(1) = 387.33$, $p < 0.001$. Separate baseline models for each timescale yielded intercepts that were significantly different from zero [cross-event, t (36.60) = 3.236, $p = 0.005$, within event, t (47.60) = −10.6, $p < 0.001$],[1] suggesting a significant difference between song and speech for each timescale. Most important, the mean score for the within-event timescale fell significantly below zero indicating the reversal of the song advantage (i.e., a speech advantage), whereas the cross-event score was significantly greater than zero.

TABLE. 1. Descriptive statistics for analyses of imitation and properties of targets.

		Timescale			
		Cross-event		Within-event	
Focus	Domain	*M*	*SD*	*M*	*SD*
Proximity[a]	Song	137.60	97.01	138.93	97.66
	Speech	222.34	77.60	203.74	79.39
	Song advantage[d]	79.80	59.12	54.84	70.32
Patterning[b]	Song	0.80	0.08	0.20	0.11
	Speech	0.73	0.14	0.54	0.12
	Song advantage	0.06	0.12	**−0.34**	0.16
SD target[c]	Song	228.626	57.345	65.645	35.541
	Speech	357.602	134.652	163.682	59.548
	Song advantage	135.925	79.417	88.685	49.294

[a]Mean absolute deviation of imitated from target pitch, in cents.
[b]Mean Pearson correlation between target and imitated f_0.
[c]Mean standard deviation of f_0 in target sequences.
[d]Pairwise differences across matched speech and song items for a given item and participant (or just item, for SD target). All differences show a song advantage except one value, highlighted in bold. All differences are significant (two-tailed, $p < 0.05$).

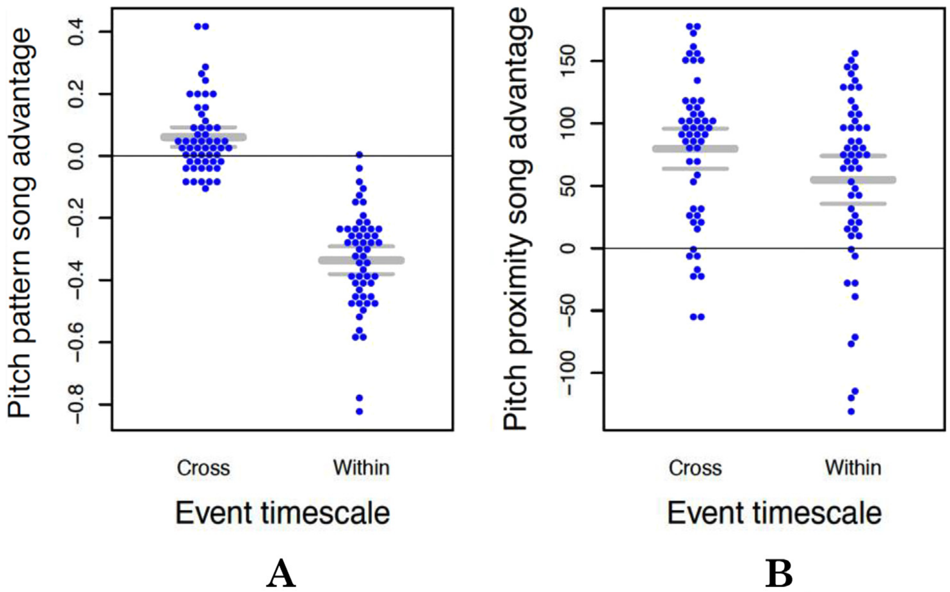

Fig. 3. Swarm plots for the song advantage as a function of timescale for pitch patterning (A) and pitch proximity (B). In each panel, bold central gray lines represent mean scores, surrounding gray lines represent 95% confidence intervals, and blue dots represent means for individual participants. Units for pitch patterning (A) are Pearson correlation coefficients, and units for pitch proximity (B) are cents. Terms in differences are arranged so that positive values always indicate an advantage for imitating song over speech; horizontal gridlines in each panel separate values exhibiting a song advantage (positive) versus a speech advantage (negative).

Likewise, the introduction of Event Timescale improved model fit to the pitch proximity measure [Fig. 3(B)], $\chi^2(1) = 10.75$, $p = 0.001$, and baseline models within each timescale yielded significantly non-zero intercepts [cross-event, $t(31.54) = 7.86$, $p < 0.001$, within event, $t(28.03) = 7.21$, $p < 0.001$], suggesting significant differences between speech and song within each timescale. In contrast to the pitch patterning measure, the mean deviation score was significantly greater than zero for each timescale, suggesting the presence of a song advantage in each case. The significant effect of Event Timescale simply reflects a greater magnitude of the advantage when assessed across events than when assessed within events.

One possible source of the song advantage reversal may lie in the structure of stimuli. Whereas sung pitches tend to be fairly stable, considerably more variability in f_0 can be found in spoken syllables. This trend is apparent in the present target stimuli. Table 1 shows variability in f_0 across events (variability across an entire sequence) as well as variability within each event, then averaged across events in a sequence. Importantly, within-event variability is nearly twice as high for speech than for song stimuli. Thus, one may question whether the low variability in song stimuli was not easily detectable by participants. I addressed this issue by adding the difference in target f_0 variability for each item as a covariate in the baseline model. The addition of Event Timescale still improved model fit for both pitch patterning, $\chi^2(1) = 139.59$, $p < 0.001$, and pitch deviation, $\chi^2(1) = 6.29$, $p = 0.012$. The reversal in the song advantage for the imitation of pitch-patterning within events does not appear to be dependent on different within-event f_0 variability across speech and song stimuli.

4. Discussion

The analysis reported reveals a reversal of the previously reported song-advantage in vocal pitch imitation [e.g., Mantell and Pfordresher (2013)]. Using combined data across comparable conditions from two previously published studies, analyses of pitch imitation accuracy replicated the song advantage with respect to the proximity of imitated to target pitch, as well as pitch patterning (i.e., correlation of imitated to target pitch over time) across an entire sequence. However, the song advantage reversed for measures of pitch patterning within events, leading to a speech advantage.

This result runs counter to the common assumption (supported by previous results) that pitch is perceived more accurately and precisely in the domain of music than for speech (Patel, 2011, 2014; Peretz and Coltheart, 2003; Zatorre and Baum, 2012). In contrast, the present results are more consistent with models proposing a trade-off between temporal and spectral precision across domains (Poeppel, 2003; Zatorre and Gandour, 2008). However, those models do not make explicit predictions for pitch perception at different timescales, or for production. Thus, the current results suggest an important revision to our understanding of pitch imitation as reflecting a multiscale planning process in which different timescales may be prioritized selectively in the process of transforming perception to action.

One subtle difference between the present analyses and previous work is that a song advantage was reported here for analyses of pitch patterning at a large timescale, whereas each individual study used here reported a null effect for this measure. It seems clear that this difference reflects statistical power, which was increased in the present paper by combining earlier data sets. Ultimately, the advantage for sung pitch in this measure is weaker than for measures of pitch proximity, which likely reflects the greater importance of absolute pitch in song versus speech [for further discussion, see Mantell and Pfordresher (2013)].

The use of a different analysis technique revealed what may be an important difference in the role of pitch across domains. Whereas sung pitch may convey information that is most significant for communication at larger timescales that span successive events, spoken pitch may convey highly significant information at smaller timescales within events (syllables). I propose that this difference reflects the use of pitch to convey meaning within each domain. Musical meaning is thought to reflect patterns of tension and release across musical events (Meyer, 1957), with fluctuations within events reflecting embellishments [e.g., vibrato, and "scoops," Larrouy-Maestri and Pfordresher (2018)] or difficulties of pitch control that are not found in in other forms of pitch-matching (Hutchins and Peretz, 2012). Many of these fluctuations in song are rapid and are found at the beginnings and ends of tones. By contrast, fluctuations of pitch within spoken syllables are considerably meaningful, in conveying focus, use of sarcasm, and question/answer distinctions in English, and may occur more gradually over time. Within-event pitch fluctuations are even more critical within tone languages. In speech, transitions in pitch across events may be comparatively less important, or at least less important than comparable transitions within music.

Acknowledgments

This work was funded in part by NSF Grant No. BCS-1848930. The author thank Ani Patel, Nicole Coleman, Tim Pruitt, Yan Chen, and an anonymous reviewer for helpful comments on an earlier version of this manuscript. The author thank Andres Buxo-Lugo and Cassandra Jacobs for guidance in applying linear mixed-effects modeling to the data.

References and links

[1]T-tests and significance values generated using the LmerTest package, with Satterthwaite's degrees of freedom method.

Albouy, P., Benjamin, L., Morillon, B., and Zatorre, R. J. (**2020**). "Distinct sensitivity to spectrotemporal modulation supports brain asymmetry for speech and melody," Science **367**, 1043–1047.

Baayen, R. H., Davidson, D. J., and Bates, D. M. (**2008**). "Mixed-effects modeling with crossed random effects for subjects and items," J. Mem. Lang. **59**, 390–412.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (**2015**). "Fitting linear mixed-effects models using lme4," J. Stat. Softw. **67**, 1–48.

Boersma, P., and Weenink, D. (**2013**). "PRAAT: Doing phonetics by computer" [computer program], version 5.4.09, http://www.praat.org/ (Last viewed 25 August 2014).

Ding, N., Patel, A. D., Chen, L., Butler, H., Luo, C., and Poeppel, D. (**2017**). "Temporal modulations in speech and music," Neurosci. Biobehav. Rev. **81**, 181–187.

Hutchins, S., and Peretz, I. (**2012**). "A frog in your throat or in your ear? Searching for the causes of poor singing," J. Exp. Psychol.: Gen. **141**, 76–97.

Kraus, N., and Chandrasekaran, B. (**2010**). "Music training for the development of auditory skills," Nat. Rev. Neurosci. **11**, 599–605.

Larrouy-Maestri, P., and Pfordresher, P. Q. (**2018**). "Pitch perception in music: Do scoops matter?," J. Exp. Psychol.: Hum. Percept. Perform. **44**, 1523–1541.

Liu, F., Jiang, C., Pfordresher, P. Q., Mantell, J. T., Xu, X., Yang, Y., and Stewart, L. (**2013**). "Individuals with congenital amusia imitate pitches more accurately in singing than in speaking: Implications for music and language processing," Atten. Percept. Psychophys. **75**, 1783–1798.

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

Meyer, L. B. (**1957**). *Emotion and Meaning in Music* (University of Chicago Press, Chicago).

Moreno, S., and Bidelman, G. M. (**2014**). "Examining neural plasticity and cognitive benefit through the unique lens of musical training," Hear. Res. **308**, 84–97.

Patel, A. D. (**2011**). "Why would musical training benefit the neural encoding of speech? The OPERA hypothesis," Front. Psychol. **2**, 142.

Patel, A. D. (**2014**). "Can nonlinguistic musical training change the way the brain processes speech? The expanded OPERA hypothesis," Hear. Res. **308**, 98–108.

Peretz, I., and Coltheart, M. (**2003**). "Modularity of music processing," Nat. Neurosci. **6**, 688–691.

Pfordresher, P. Q., Mantell, J. T., and Pruitt, T. A. (**2021**). "Effects of intention in the imitation of sung and spoken pitch," Psychol. Res. (published online).

Poeppel, D. (**2003**). "The analysis of speech in different temporal integration windows: Cerebral lateralization as 'asymmetric sampling in time,'" Speech Commun. **41**, 245–255.

Sares, A. G., Foster, N. E. V., Allen, K., and Hyde, K. L. (**2018**). "Pitch and time processing in speech and tones: The effects of musical training and attention," J. Speech Lang. Hear. Res. **61**, 496–509.

Wang, L., Pfordresher, P. Q., Jiang, C., and Liu, F. (**2021**). "Individuals with autism spectrum disorder are impaired in absolute but not relative pitch and duration matching in speech and song imitation," Autism Res. **14**, 2355–2372.

Wisniewski, M. G., Mantell, J. T., and Pfordresher, P. Q. (**2013**). "Transfer effects in the vocal imitation of speech and song," Psychomusicol.: Music Mind Brain **23**, 82–99.
Zatorre, R. J., and Baum, S. R. (**2012**). "Musical Melody and Speech Intonation: Singing a different tune," PLoS Biol **10**, e1001372.
Zatorre, R. J., Belin, P., and Penhune, V. B. (**2002**). "Structure and function of auditory cortex: Music and speech," Trends Cogn. Sci. **6**, 37–46.
Zatorre, R. J., and Gandour, J. T. (**2008**). "Neural specializations for speech and pitch: Moving beyond the dichotomies," Philos. Trans. R. Soc. B: Biol. Sci. **363**, 1087–1104.