The effect of musical training and language background on vocal imitation of pitch in speech and song

The effect of musical training and language background on vocal imitation of pitch in speech and song

Chihiro Honda[1], Tim Pruitt[1], Emma Greenspon[2], Fang Liu[3], & Peter Q. Pfordresher[1]

[1]University at Buffalo, the State University of New York, NY, USA

[2]Monmouth University, NJ, USA

[3]University of Reading, UK

Abstract

Vocal imitation plays a critical function in the development and use of both language and music. Previous studies have reported more accurate imitation for sung pitch than spoken pitch, which might be attributed to the structural differences in acoustic signals and/or the distinct mental representations of pitch patterns across speech and music. The current study investigates the interaction between bottom-up (i.e., acoustic structure) and top-down (i.e., participants' language and musical background) factors on pitch imitation by comparing speech and song imitation accuracy across four groups: English and Mandarin speakers with or without musical training. Participants imitated pitch sequences that were characteristic of either song or speech, derived from pitch patterns in English and Mandarin spoken sentences. Overall, song imitation was more accurate than speech imitation, and this advantage was larger for English than Mandarin pitch sequences, regardless of participants' musical and language experiences. This effect likely reflects the perceptual salience of linguistic tones in Mandarin relative to English speech. Music and language knowledge were associated with optimal imitation of different acoustic features. Musicians were more accurate in matching absolute pitch across syllables and musical notes compared to non-musicians. By contrast, Mandarin speakers were more accurate at imitating fine-grained changes within and across pitch events compared to English speakers. These results suggest that different top-down factors (i.e., language and musical background) influence pitch imitation ability for different dimensions of bottom-up features (i.e., absolute pitch and relative pitch patterns).

Keywords: vocal imitation, pitch imitation, tone language, musical training

Public Significance Statement

Results of this study revealed that pitch imitation ability is influenced by one's language and musical background as well as the characteristics of acoustic stimuli. Musical training may improve the ability to match absolute pitch whereas experience with a tone language may enhance the ability to imitate relative pitch.

The effect of musical training and language background on vocal imitation of pitch in speech and song

The human ability to vocally imitate sounds is crucial for vocal learning, such as learning to sing an unfamiliar song or learning to speak an unfamiliar language. Vocal imitation is defined as an action that attempts to match the acoustic characteristics of sound signals using the vocal motor system (Mercado et al., 2014). Previous work on this topic has focused on the imitation of pitch in order to compare the ability in speech and music domains. Several studies have found an advantage for imitating sung pitch over spoken pitch (F. Liu et al., 2013; Mantell & Pfordresher, 2013; Pfordresher et al., 2022; Wang et al., 2021; Wisniewski et al., 2013). This advantage may be due to the structural differences in acoustic signals of sung and spoken sequences (i.e., stable vs. variable pitch; Pfordresher et al., 2022; Stegemöller et al., 2008) or the mental representations of pitch patterns that facilitate tonal encoding (Peretz & Coltheart, 2003). In fact, there are individual differences in pitch imitation ability both within and across music and language domains. First, whereas most adult individuals are typically able to match musical notes within a semitone, a small number of the population are generally inaccurate at matching pitches (Pfordresher & Demorest, 2021; Berkowska & Dalla Bella, 2013). Second, inaccurate imitators tend to show deficits in both sung and spoken pitch imitation (F. Liu et al., 2013; Mantell & Pfordresher, 2013; Pfordresher et al., 2022; Wisniewski et al., 2013). Finally, experiences such as singing lessons can affect singing accuracy (e.g., Demorest et al., 2018). Thus, both acoustic structures and individual experience with pitch seem to affect vocal imitation accuracy.

Top-down vs. Bottom-up Processing

In the current study, we investigated the effects of acoustic structures and long-term experience on pitch imitation ability by using the concept of bottom-up and top-down processing. We define these factors following Bregman (1990), who proposed that the formation of auditory representations consists of two kinds of systems, primitive (i.e., bottom-up) and schema-driven (i.e., top-down) processes. The primitive processes are driven by incoming acoustic information whereas schema-driven processes are activated by matching pre-existing knowledge of familiar sound patterns (i.e., schemas) to primitive data. These processes are guided by underlying neural mechanisms. Specifically, auditory bottom-up factors involve stimulus-based sensory information (e.g., speech and song stimuli) ascending from the cochlea to the cortex whereas top-down factors provide experience-based feedback pathways (e.g., tonal knowledge) descending from the auditory cortex to the brainstem (Kraus & Chandrasekaran, 2010), or from higher cortical areas to lower cortical areas (Nahum et al., 2008) to facilitate processing of incoming sounds. This suggests that pitch processing can be affected by both the structures of auditory information (i.e., bottom-up factor) and the schemas that the listeners have formed in their long-term experience (i.e., top-down factor).

Consistent with Kraus and Chandrasekaran (2010), previous research suggests that the neural activity involved in auditory processing differs depending on bottom-up factors (e.g., music vs. speech; e.g., Zatorre et al., 2002) and top-down factors (e.g., musical training; Wong et al., 2007). There are behavioral differences in auditory tasks based on bottom-up and top-down factors. First, vocal imitation accuracy is affected by bottom-up factors (e.g., pitch structures), where people generally imitate stable pitches (e.g., melodies) more accurately than variable pitches (e.g., speech; Mantel & Pfordresher, 2013). Second, several studies have shown behavioral improvements by long-term experience, such as musical and language training; for example, six-month musical training improved pitch discrimination in both language and music in dyslexic children (Besson et al., 2007). Longitudinal evidence further suggests that musical training is associated with both behavioral and neural changes. Habibi and colleagues (2018) reported that children who received musical training exhibited better auditory performance than those in control groups and neural changes in related brain regions. Thus, neural and behavioral evidence supports both acoustic structures (i.e., bottom-up factors) and long-term experience (i.e., top-down factors) can influence auditory processing.

Enhancement of top-down processing in one domain (e.g., musical training) may facilitate auditory processing in a different domain (e.g., speech) via *transfer of training*. Whereas some studies suggest that mechanisms involved in music and language processing are based on separate modules (Peretz & Coltheart, 2003), others argue that the two domains share overlapping cognitive resources (Patel, 1998; Van de Cavey & Hartsuiker, 2016). Patel (2011; 2014) proposed the OPERA (overlap, precision, emotion, repetition, and attention) hypothesis which states that musical training benefits speech processing under certain conditions. According to this hypothesis, transfer of musical training occurs when brain networks that process music overlap with those for processing speech, when music requires higher precision of encoding acoustic features than speech, and when musical activities are associated with strong emotion and require focused attention. Consistent with this hypothesis, evidence suggests that musical training facilitates speech processing of lexical tones (Lee & Hung, 2008) and leads to higher sensitivity to prosody (Thompson et al., 2004). However, it is important to note that there are mixed results when it comes to transfer effects from musical training to other cognitive abilities. For instance, a recent study found no effect of musical training on speech perception in background noise (Madsen et al., 2019). As the authors observed, exposure to speech in noise is a common experience for people regardless of musical background. Musical training may therefore not add much to the effects of exposure; people regardless of musical background have exposure to situations in which they hear and comprehend speech in noisy environment, and musical training does not provide any additional benefits to the ability to hear speech in noise. Taken together, musical training may enhance speech processing by strengthening overlapped neural networks, but such transfer effects might be limited to certain closely-associated abilities, such as pitch processing.

Although the OPERA hypothesis mainly focuses on beneficial transfer from the music domain to the speech domain, the opposite direction of transfer may occur; that is, certain language background may also influence musical ability. For example, tonal languages, such as Mandarin and Cantonese, utilize pitch patterns assigned in lexicons called *lexical tones*, to distinguish word meanings (Yip, 2002). As found for research on transfer effects from music to language, previous studies have also found mixed results for the beneficial transfer of language ability to the music domain. Whereas some behavioral studies have found no effects of tonal language background on certain musical tasks (e.g., musical pitch perception; Bidelman et al., 2011), others have found an advantage of tonal language background for auditory processing (e.g., simple pitch discrimination; Guiliano et al., 2011). For instance, Bidelman and colleagues (2013) showed that tonal language speakers (i.e., Cantonese speakers) outperformed English speaking non-musicians on fundamental frequency difference limens, pitch memory, and musical melody discrimination (differing by ½ semitones) tasks, suggesting the benefits of tonal language background on pitch perception ability. Moreover, another study suggests that tonal language speakers are better at discriminating musical intervals and imitating melodic intervals via singing than non-tonal language speakers (Pfordresher & Brown, 2009). A recent meta-analysis and large-sample online study further suggests an overall benefit for musical pitch processing among tone language speakers (J. Liu et al., 2023). Therefore, long-term experience with a tonal language may also provide a beneficial effect on auditory processing of certain musical features and for certain tasks.

The present study further investigates the interaction of bottom-up and top-down factors on pitch imitation ability; specifically, we ask whether pitch imitation accuracy is affected by bottom-up acoustic structures that represent the pitch/time trajectories present in music and speech (hereby referred to simply as music and speech for brevity) as well as participants' language and musical background (top-down). The current study utilized the vocal imitation paradigm developed by Mantell and Pfordresher (2013) in which participants listened to and vocally imitated phonetically neutral pitch contours (the holistic trajectory of pitch change within a sequence) from either English sentences or from melodies based on the pitches associated with spoken syllables. Whereas Mantell and Pfordresher (2013) investigated the structural differences in acoustic sequences (speech vs song) based on English speech only, the current study extends the scope by including stimuli based on Mandarin speech. In addition, the current study further explores the role of top-down processing by recruiting participants from different background (i.e., musicians vs non-musicians and English vs Mandarin speakers). Finally, the present study focuses on simplified versions of the original recordings that include only the evolution of pitch (f_0) over time, to avoid complications associated with having participants imitate phonetic features of an unfamiliar language.

In the current study, we utilized sentences in a non-tonal language (i.e., English) and a tonal language (i.e., Mandarin) to compare the structural differences within the speech domain. Whereas Mandarin speech stimuli were based on four canonical tone categories used for each syllable (level, rising, falling-rising, and falling), English speech does not use such tonal categories. In addition, Mandarin speakers and English speakers were recruited for the current study for the comparison between tonal language and non-tonal language background, and both speaker groups were further classified into musician or non-musician based on total years of musical training. Following Mantell and Pfordresher's (2013) results, we hypothesized that participants would imitate musical sequences more accurately than speech sequences. Moreover, based on previous research on transfer effects between music and language domains, we also hypothesized that those with musical training and/or Mandarin language background would show enhanced imitation ability in both music and language domains.

Method

Participants

One hundred and twenty-seven participants (M_{age} = 20.59, SD_{age} = 3.41; 55.12% female) from the University at Buffalo community were recruited via online posting associated with the Introduction to Psychology participant pool (SONA system), flyers posted around campus, or emails sent to student mailing lists. Participants either received a course credit or a $10 gift card for their participation. Participants' native language (the first language they have acquired) was either English (N = 51) or Mandarin (N = 76). While all Mandarin speakers have learned English language, only a few English speakers (N = 4) have had exposure to a tonal language.

Table 1 describes the means and standard deviations of age, years of musical training and experience, and pitch discrimination score for each group. Musical training is defined as formal training (i.e., private lessons) of any musical instrument whereas musical experience is informal experience with playing any musical instrument. We used years of musical training to classify musicians and non-musicians based on previous studies (e.g., Pfordresher et al., 2020). Participants in both language groups with at least three years of formal musical training were classified as musicians (N = 44) or as non-musicians (N = 83) if they have less than three years of musical training. Analysis of variance yielded a main effect of language background on musical experience, $F(1, 123) = 4.53$, $p = .035$, $\eta^2_p = .04$, suggesting that English speakers had more years of musical experience, but there was no effect of language background on musical training ($p = .41$). Finally, there were no effects of grouping variables on pitch discrimination (all $p > .1$; see below for details).

Participants were given the option to participate virtually or in-person (see *Procedure*); 85 participants completed the experiment virtually (52 Mandarin; 42 female; 28 musicians) while 42 participants completed the experiment in-person (24 Mandarin; 28 female; 16 musicians)[1]. Analyses comparing participants' performance in these experimental settings yielded no differences for the variables reported here. Seven additional participants (not included in the total counts above) were excluded from the analyses due to technical issues during the experiment or failure to follow the experimental instruction.

Table 1
Means (standard deviations) of age, years of formal musical training and musical experience, and average pitch discrimination scores for each group.

Language/Musical background group	N	Female N / Male N	Age	Musical Training	Musical Experience	Pitch Discrimination
English/Musician	21	12 / 9	19.29 (1.95)	7.21 (3.18)	8.6 (3.61)	96.76% (3.66)
Mandarin/Musician	23	14 / 9	23.43 (4.65)	6.13 (4.35)	7.04 (4.41)	89.30% (20.15)
English/Non-musician	30	17 / 13	18.77 (0.86)	0.23 (0.63)	2.30 (2.82)	88.87% (12.66)
Mandarin/Non-musician	53	27 / 26	20.91 (3.33)	0.31 (0.63)	1.20 (2.75)	88.38% (14.25)

Our sampling strategy was based on a power analysis using estimated effect size for the previously documented advantage for imitating song versus speech, based on absolute pitch deviation scores (see later discussion). Results reported in Pfordresher (2022) for this effect yielded estimated Cohen's d = .94, one tailed (a large effect), using G-power (Faul et al., 2007).

[1] A virtual setting was offered to accommodate the restrictions on physical contacts during the COVID-19.

Based on this estimate, detecting a statistically significant song advantage at a power of .95 within a single group would require a sample size of n = 14. We applied this rubric for each treatment group in the current design (k = 4 groups), and intentionally set a higher target of n = 20 per group given that other effects of interest may be of lower power than the song advantage. During sampling, English and Mandarin-speaking participants were allowed to sign up without precondition with respect to musical background group, for purposes of allowing equal access for Introduction to Psychology research credit, and so sampling continued until this number or more was reached per group. Analyses used to address possible issues resulting from unequal sample sizes, none of which yielded concerns for the present results, are reported in Appendix C.

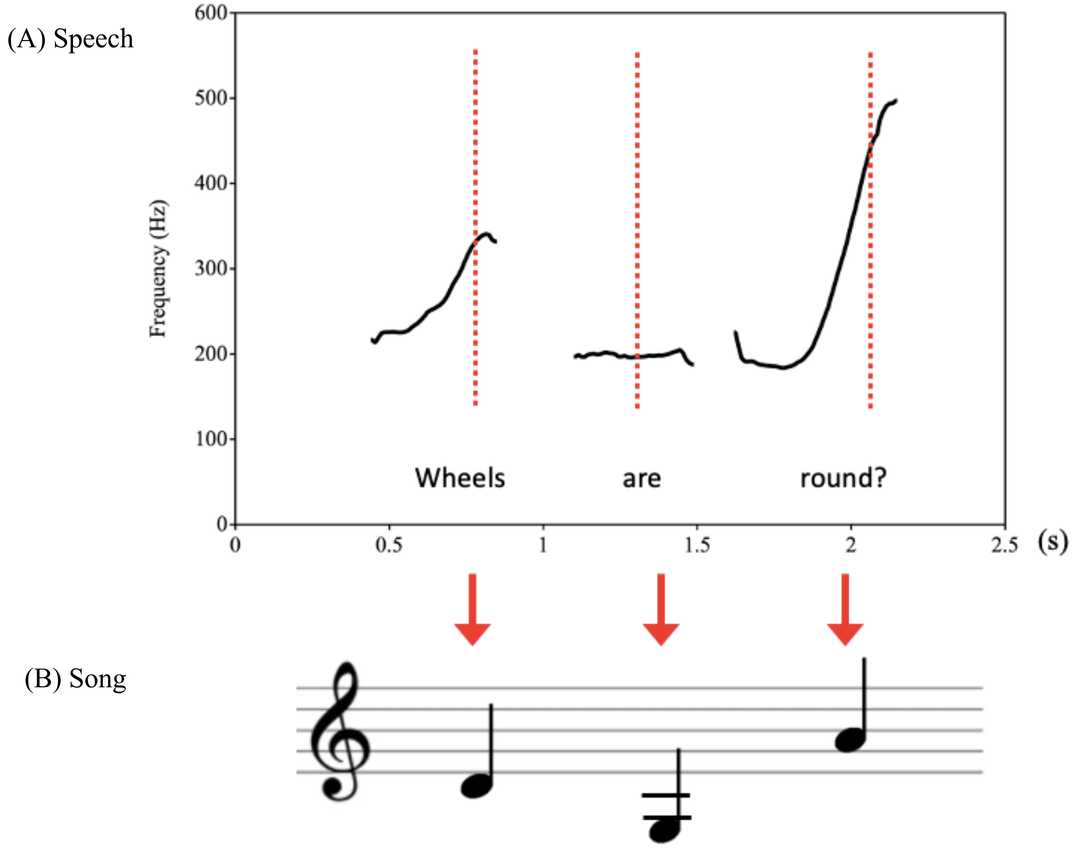
Target stimuli

Speech stimuli were 48 short sentences consisting of 12 original texts translated in two target languages (English and Mandarin; see *Appendix A*) produced with two pitch contours (statement and question, which serve to vary pitch patterns in the phrase). Each sentence contained three to five syllables, and the place of emphasis on words varied across sentences. The English and Mandarin sentences had different meanings but shared the same number of syllables as well as similar pitch contours at word- and sentence-levels (see *Appendix A*). The main difference between these two types of language stimuli was that Mandarin sentences had systematic pitch changes at the syllable-level which consisted of four tones (level, rising, falling-rising, and falling) whereas the pitch changes in the English syllables were not based on such lexical tones. Sentences used for female participants were produced by a female native speaker of each language (one English native speaker, one Mandarin native speaker) and sentences used for male participants were produced by a male native speaker of each language (one English native speaker, one Mandarin native speaker). The recordings of produced speech were converted to a phonetically neutral pitch trace using the "hum" function of Praat (Boersma & Weenik, 2013) which can be used to present a pitch trace in a voice-like timbre. The resulting stimulus sounds like a neutral vowel and has a uniform timbre (i.e., filtered five formants), but precisely replicates changes in pitch over the course of the utterance. These phonetically neutral sequences were used for the experiment.

The song stimuli were 48 melodies created from the pitch information in the speech stimuli described above. First, we identified the point in the syllable at which the *f*0 best matched the perceived global pitch of the syllable based on subjective evaluation by one of the authors (CH). – generally speaking, the highest pitch within each syllable was selected. Then, the *f*0 at each identified point was assigned to the closest corresponding diatonic pitch in the G major scale. Figure 1 illustrates an example of a speech stimulus produced in English by a native female speaker and its corresponding song stimulus, where the representative pitch of each syllable was identified and sequentially assigned to D4, G3, and A4. Therefore, the melodic contour in each song stimulus approximately matched the pitch contour of the corresponding speech stimulus, but with stable pitches. Pre-recorded notes produced by male and female vocalists (used for the Seattle Singing Accuracy Protocol; see Pfordresher & Demorest, 2020) were concatenated to create melodies, and each note in the melodies carried the same duration (750 ms) to create isochronous melodies with a steady rhythm. Finally, these melodies were converted to hums in the same way as for the speech stimuli, and the phonetically neutral melodic sequences were used for the experiments. No significant differences in overall pitch height or pitch variability were present across the 4 stimulus categories used in this study, as detailed in Appendix B.

To equate temporal characteristics of speech and song stimuli, we adjusted the durations of speech stimuli in the following way. The duration of each speech stimulus was altered based on the number of syllables to match its song counterpart discussed in the next section (3-syllable sentence = 2250 ms, 4-syllable sentence = 3000 ms, 5-syllable sentence = 3750 ms of total duration). This alteration resulted in an adjustment of 127.91% for durations of the original speech recordings on average (3-syllable sentence = 114.70%, 4-syllable sentence = 130.68%, 5-syllable sentence = 138.34%).

Figure 1
An Example of a speech stimulus and its corresponding song stimulus.



Note: Black lines in (A) represent f_0 in an example speech stimulus, “Wheels are round?”. The dotted lines indicate the points in time where the pitch was extracted. Musical notes in (B) shows the corresponding song stimulus with red arrows indicating mapping of syllables to notes.

To ensure the validity of our stimuli (e.g., song stimuli are perceived as songs), we conducted a rating task prior to collecting data for the main experiment using different participants from those reported above. Fifteen native English speakers and nine native Mandarin speakers rated whether the target stimulus sounds like speech or song using a 7-point Likert scale from 1 (“clearly speech”) to 7 (“clearly song”), with 4 indicating “neutral”. Overall, participants rated song stimuli (English song: *M* = 6.97, *SD* = 0.88; Mandarin song: *M* = 6.93, *SD* = 0.92) higher than speech stimuli (English speech: *M* = 2.05, *SD* = 0.88; Mandarin speech: *M* = 2.63, *SD* = 1.00) in both stimulus language types. Table 2 shows the means and standard deviations of the ratings on each stimulus category for each language background group. Both English and Mandarin speakers rated song stimuli significantly higher than speech stimuli in each stimulus language (i.e., English song > English speech and Mandarin song > Mandarin speech), *p* < .001 for all pairs. Interestingly, English speakers rated Mandarin speech significantly higher than English speech, *t*(26.31) = 3.38, *p* < .001, while Mandarin speakers did not rate these two types of speech stimuli differently, *p* = .38. This might reflect the participants’ language background – English speakers were not familiar with Mandarin speech while Mandarin speakers are familiar with both Mandarin and English speech. Therefore, our English and Mandarin speech stimuli might be perceived differently (i.e., Mandarin speech as more “song-like” than English speech) by the English background group whereas song stimuli in both kinds of languages are perceived as song by both language background groups.

Table 2
Means (standard deviations) of perception ratings on stimuli for each language group.

	N	English Speech	Mandarin Speech	English Song	Mandarin Song
English Speakers	15	1.84 (0.61)	2.52 (0.47)	6.63 (0.33)	6.56 (0.42)
Mandarin Speakers	9	1.58 (0.58)	1.88 (0.82)	6.54 (0.51)	6.55 (0.44)

Tasks

Pitch Discrimination

The task consisted of ten sets of pure tone pairs; the first tone was always 500 Hz in frequency, and the second tone was one of the following: 300 Hz, 350 Hz, 400 Hz, 450 Hz, 475 Hz, 525 Hz, 550 Hz, 600 Hz, 650 Hz, and 700 Hz. Each tone was played for 1 s, and there was a 500 ms pause between tones. On each trial, participants heard the 500 Hz tone and indicated whether the second tone was higher or lower in pitch than the first tone. Each pair was presented five times (50 trials total), and the order of the pairs was randomized.

Vocal Warmups

Vocal warmups consisted of reading aloud and singing. Reading materials comprised a short English passage "The Eagle" by Alfred Lord Tennyson (for both English and Mandarin speakers) and a Mandarin passage "静夜思" by Li Bai (for Mandarin speakers only). The reading task was used to ensure that participants were able to read fluently in their native language; no participants were disqualified based on their reading ability. In the singing task, participants were asked to sing a note that they found comfortable to sing and then were asked to sing the highest and lowest notes they could sing. All sung pitches in the warm-up singing task were held for at least two seconds each.

Single-Pitch Matching

Participants listened to a single note and imitated the pitch using the syllable "doo". The same pre-recorded notes used for the experimental stimuli were used for this task. For each trial, male participants imitated one of five notes (C3, D3, E3, F3, and G3) produced by a male vocalist twice, and female participants imitated one of five notes (C4, D4, E4, F4, and G4) produced by a female vocalist twice.

Experimental Pitch Imitation trials

On each trial, participants listened to one of the target stimuli and then imitated the stimulus as accurately as possible with their voice. Male participants were presented the stimuli recorded by male speakers, and female participants were presented the stimuli recorded by female speakers. The pitch imitation task consisted of 96 trials (with 48 speech and 48 song stimuli). Each participant was assigned to one of two randomized orders.

Apparatus

For the in-person setting, participants were tested individually in a sound-attenuated booth (WhisperRoom Inc.). Stimuli were generated using in-house Matlab programs (Mathworks, Natick, MA) interfacing with a USB audio interface (Focusrite Scarlett 2i2). Participants listened to stimuli using over-ear headphones (Senheiser HD280 Pro). Vocal productions were recorded using a Shure PG58 and recorded digitally using the same USB interface and Matlab programs used to present stimuli. The texts for the warm-up as well as the choices for the discrimination task were presented on a monitor (Dell Inc.). Participants' numerical responses for the discrimination task were recorded via a keypad (Targus).

For the online setting, the initial instructions and questionnaires were given via Zoom (Zoom Video Communications), and the experimental tasks were given on a web-based platform, FindingFive (FindingFive Corporation). Participants were instructed to take the experiment in a quiet environment and to use their own headphones and external microphone if they were available.

Procedure

The experiments were conducted in in-person and online settings. This study was approved by the Institution Review Board, University at Buffalo, SUNY.

In-person setting

The in-person experiment was conducted following the Health and Safety guidance during the COVID-19 pandemic from the University at Buffalo, SUNY. Participants and the experimenter both wore a facial mask and maintained at least six-foot distance from each other. Participants then completed experimental tasks in the sound attenuated booth. Instructions for each task were given in participants' native language (English or Mandarin). After the experimental tasks, participants completed a questionnaire about their language and musical background.

Online setting

Participants received instructions from the experimenter live via Zoom (Zoom Video Communications). Participants then completed the experimental tasks on FindingFive (FindingFive Corporation). The Zoom session continued so that the experimenter could monitor the participant's progress and compliance; however, the participant and experimenter's camera remained off to maintain privacy. Instructions for each trial were written in participants' native language and were projected on a screen. Then, participants were given the same questionnaire by the experimenter via Zoom.

Data Analyses

f_0 values were extracted from each recording using Praat at least a month after the session. The extractions were performed in Praat using the default settings for its autocorrelation algorithm, at a sample rate of one f_0 value every 10 ms (100 Hz). Research assistants evaluated each participant recording for octave errors and pitch artifacts by comparing the audio recording of the participant with a synthetic rendering of the extracted f_0 values using the Praat "hum" timbre. Extraction errors were defined as instances in which some portion of the perceived pitch from the extraction did not match the perceived pitch in the original recording. Causes of such artifacts included octave errors in the extraction algorithm, or occasions in which pitch was attributed to a non-pitched sound (e.g., a participant coughs, or there is an environmental noise during recordings done at home). Corrections were performed manually in Praat using one or more parameters (pitch floor, pitch ceiling, silence threshold, voicing threshold, octave cost, octave-jump cost, and voicing/unvoiced cost), until the vocal pitch matched across versions. If no match could be obtained the trial was discarded; this occurred for 5.29% of all trials. At no point in time was the original target stimulus used as a reference in this process.

The extracted f_0 as well as target f_0 values were converted from hertz (Hz) to cents with a baseline of 98 hertz (around G2) for males or 215 hertz (around A3) for females (cents = 1200 x $\log_2$ (observed Hz/baseline Hz)). In-house Matlab scripts compared the f_0 values of the imitations with the target f_0 values, the duration of each target was adjusted to match the duration of the corresponding imitation by resampling the target pitch vector so that each sampled f_0 value aligned with the corresponding f_0 value from the imitation based on its relative sequential position.

For the main analyses, we focused on two measures to evaluate imitation accuracy: *pitch deviation* and *pitch correlation*. *Pitch deviation* refers to the mean absolute difference between all sampled f_0 values of the imitation and their temporally matched f_0 values from the target for a trial. This measure assessed how accurately participants match the pitch of the target. First, all f_0 values were converted from Hz to cents (100 cents = 1 semitone). Then, the absolute pitch deviation for each data point was calculated by subtracting the f_0 values of the imitation from matched f_0 values of the target and taking the absolute value of the difference. The resulting absolute pitch deviation values in each trial were then averaged, and the mean pitch deviations across trials were calculated for each individual. *Pitch correlation* measured how accurately participants imitate the patterns of relative pitch of the target over time for a given trial. A pitch correlation value was calculated by regressing all sampled f_0 values in the imitation on the corresponding f_0 values in the target for each trial, and the mean pitch correlations across trials were calculated for each individual. *Pitch correlation* is used to assess the imitation accuracy of relative pitch whereas *pitch deviations* focus on absolute pitch.

Each dependent variable was analyzed using a 2 (stimulus language: English and Mandarin stimulus) × 2 (stimulus domain: speech contour vs song) × 2 (language background: English vs Mandarin speakers) × 2 (musical background: musician vs non-musician) mixed-model Analysis of Variance (ANOVA). Stimulus language and stimulus domain were within participants factors, whereas language and musical background were between participants factors. Planned comparisons focusing on background (i.e., musicians vs non-musicians and English vs Mandarin speakers) in each domain were performed based on our theoretical predictions concerning the influence of background differences on top-down processing. The two dependent variables of interest were not normally distributed. However, transformations that yielded best approximations to normal distributions according to Q-Q plots (a log-10 transformation for pitch deviations and logistic transformation for pitch correlations) yielded the same pattern of significance as found for untransformed data. For sake of clarity, we report results from untransformed data below.

Transparency and Openness

The study was not preregistered. The data and code of this paper are available at the project's OSF page https://osf.io/9rmha/.

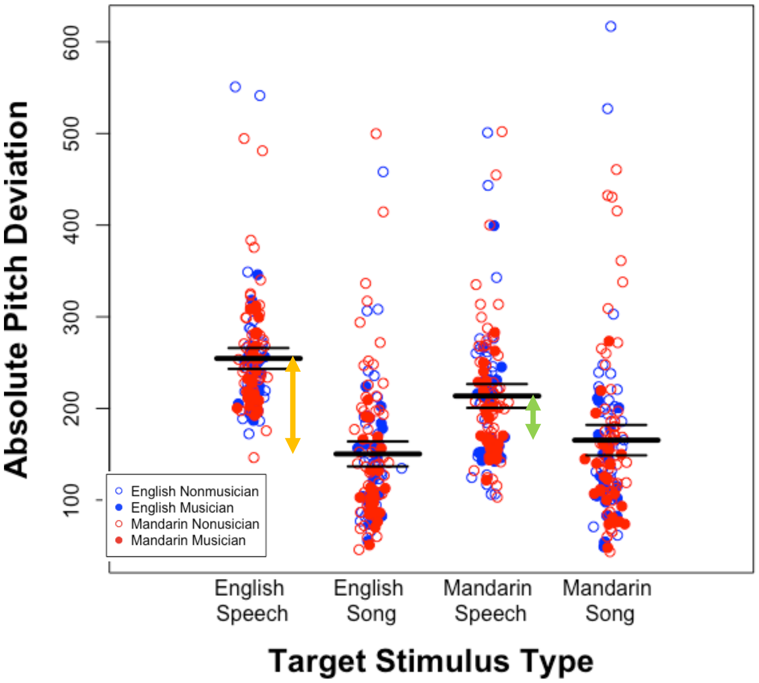
Results

Pitch Deviation Analysis for Pitch Imitation

First, we analyzed the pitch imitation data by using pitch deviation as a dependent measure. Figure 2 illustrates the average pitch deviation for each participant across all trials in a condition as well as means and 95% confidence intervals for each group. The ANOVA yielded a significant main effect of stimulus domain, $F(1, 123) = 315.58$, $p < .001$, $\eta^2_p = .72$, indicating that pitch deviation scores for song imitation ($M = 157.84$, $SD = 80.55$) were significantly lower than those for speech imitation ($M = 234.07$, $SD = 66.85$), as in Mantell and Pfordresher (2013). There also was a main effect of stimulus language, $F(1, 123) = 12.47$, $p < .001$, $\eta^2_p = .09$, indicating that pitch deviation scores for Mandarin stimuli ($M = 189.48$, $SD = 80.25$) were significantly lower than those for English stimuli ($M = 202.43$, $SD = 64.64$).

Also, there was a significant stimulus domain × stimulus language interaction, $F(1, 123) = 102.11$, $p < .001$, $\eta^2_p = .45$. To analyze this interaction, we first calculated the difference between speech and song imitation by subtracting the average pitch deviation in song imitation from the average pitch deviation in speech imitation (the positive value reflects a song advantage) for each individual and each stimulus language. The contrast between speech and song was significant within each stimulus language condition ($p < .001$ in each case). We then ran *t*-tests between the stimulus languages using the difference as a measure for the degree of song advantage. In Figure 2, the yellow arrow illustrates the song advantage for English stimuli, and the green arrow illustrates the song advantage for Mandarin stimuli. A paired *t*-test yielded a significant difference between the stimulus languages, $t(126) = 10.11$, $p < .001$, indicating a greater song advantage for English stimuli than for Mandarin stimuli. The same pattern of results was observed for both male and female participants.

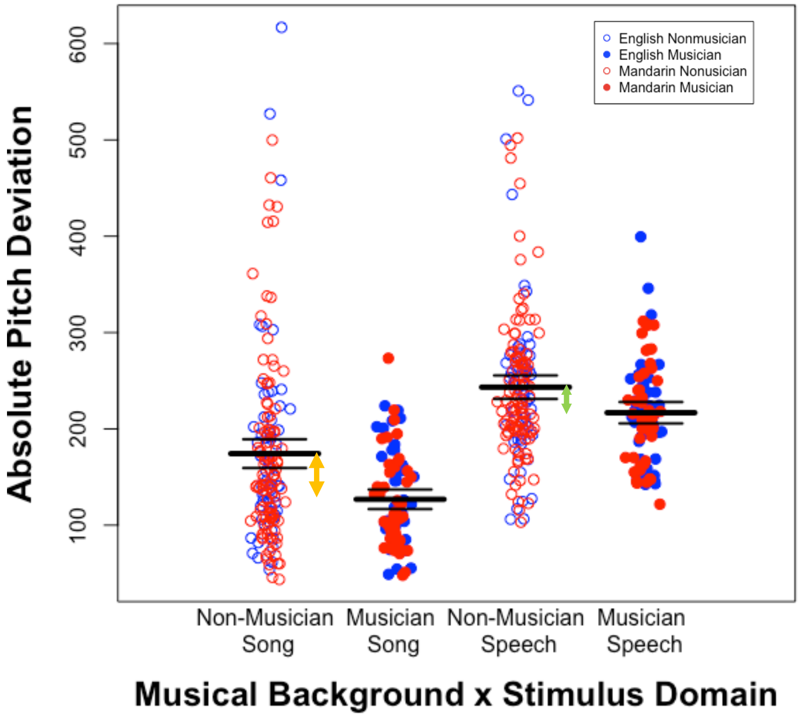
Figure 2
Pitch deviation scores between the imitation and target pitch for each stimulus type.



Note: Each dot indicates each participant's mean scores (blue = English speakers, red = Mandarin speakers, open = non-musicians, and closed = musicians). The bold lines show group means and upper and lower lines represent 95% confidence intervals. The arrows indicate song advantage (speech – song imitation accuracy) for English (yellow) and Mandarin (green) stimuli.

There was also a significant main effect of musical background, $F(1, 123) = 8.52$, $p = .004$, $\eta^2_p = .07$ and a significant musical background × stimulus domain interaction, $F(1, 123) = 5.49$, $p = .021$, $\eta^2_p = .04$. Figure 3 illustrates the means and confidence intervals of song and speech imitation for each musical background group. Pitch deviation scores were lower for participants with at least three years of musical training ($M = 171.80$, $SD = 38.09$) than for participants with less than three years of musical training ($M = 208.76$, SD = 79.13). Planned comparisons between musicians and non-musicians indicated that musicians outperformed non-musicians in both song imitation, *t*(123.66) = 4.00, *p* < .001 (the yellow arrow in Figure 3) and speech imitation, *t*(124.68) = 2.54, *p* = .012 (the green arrow in Figure 3), suggesting that the interaction reflected the magnitude of the musician advantage across domains. Finally, there was no main effect of language background and no other interactions.

Figure 3
Pitch deviation scores between imitation and target pitch for each musical background group.



Note: Each dot indicates each participant's mean scores (blue = English speakers, red = Mandarin speakers, open = non-musicians, and closed = musicians). The bold lines group means and upper and lower lines represent 95% confidence intervals. The arrows indicate advantage of musical training (non-musician – musician) for song (yellow) and speech (green) stimuli.

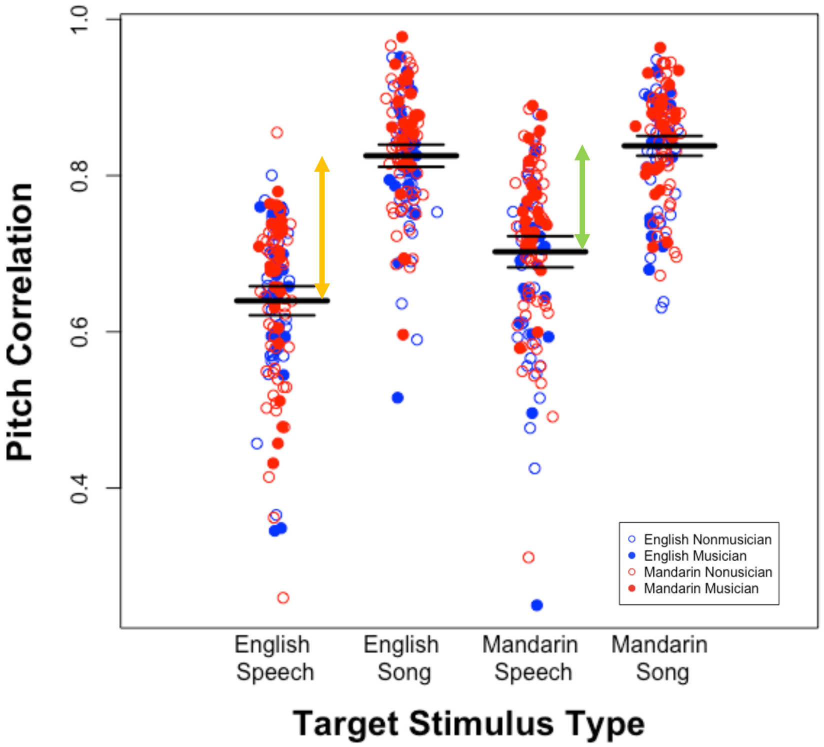
Observed effects for the production data in the main experiment follow similar patterns to the perceptual ratings of stimuli completed by an independent group of listeners (reported in *Stimuli*). As such, it is unclear to what degree the observed differences in pitch deviation scores may be reduced to differences in perceptual evaluations across a speech/song continuum. We addressed this point (brought up by an anonymous reviewer), through linear detrending based on associations between ratings and pitch deviation scores. Each participant's mean deviation score for each stimulus condition (stimulus domain × stimulus language) in the main experiment was correlated with the mean perceptual rating provided by participants in the rating task who shared the same native language with that participant. The resulting correlation across all participants and stimulus conditions was significant, $r(11545) = -.28$, $p < .001$ validating the overall association between ratings and production. Next, we removed variability in pitch deviation scores that were associated with this correlation through linear detrending, and re-ran the ANOVA on detrended pitch deviation scores. The most critical effects for the present study remained significant, namely the main effect of musical background, $F(1,123) = 8.07$, $p = .005$, $\eta^2_p = .062$, the stimulus domain × stimulus language interaction, $F(1,123) = 71.08$, $p < .001$, $\eta^2_p = .366$, and the musical background × stimulus domain interaction, $F(1,123) = 4.57$, $p = .035$, $\eta^2_p = .036$. The only effect that no longer remained significant in the detrended data was the main effect of stimulus domain (detrended $p = .63$, $\eta^2_p = .002$).

Pitch Correlation Analysis for Pitch Imitation

Next, we performed the same analysis by using pitch correlation as a dependent variable. Figure 4 illustrates the mean pitch correlation score for each participant and condition as well as the mean and 95% confidence intervals. As in the pitch deviation analyses described above, there were significant main effects of stimulus domain $F(1, 123) = 429.57$, $p < .001$, $\eta^2_p = .78$, indicating that the pitch correlations for song imitation were overall higher ($M = .83$, $SD = .07$) than that for the speech imitation ($M = .67$, $SD = .10$), as well as stimulus language, $F(1, 123) = 74.84$, $p < .001$, $\eta^2_p = .38$, indicating that the pitch correlations for Mandarin stimuli were higher ($M = .77$, $SD = .08$) than that of English stimuli ($M = .73$, $SD = .08$).

Figure 4

Pitch correlation between the imitation and target pitch.



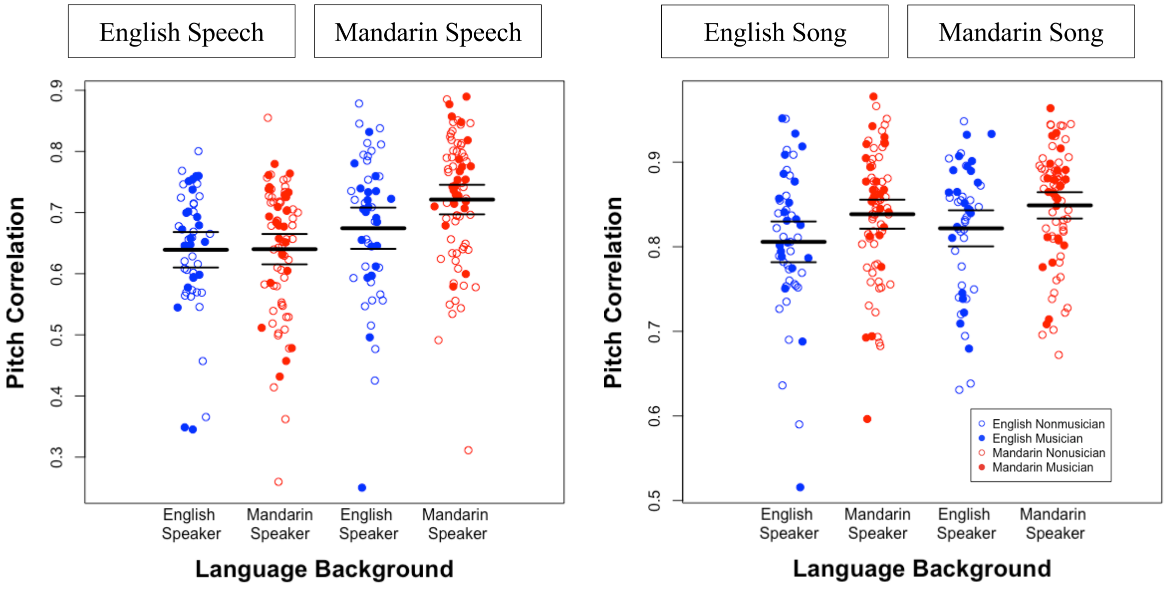
Note: Each dot indicates each participant's mean scores (blue = English speakers, red = Mandarin speakers, open = non-musicians, and closed = musicians). The bold lines show group means and upper and lower lines represent 95% confidence intervals. The arrows indicate song advantage (song – speech imitation accuracy) for English (yellow) and Mandarin (green) stimuli.

We also found a significant two-way interaction between stimulus domain and stimulus language, $F(1, 122) = 33.90$, $p < .001$, $\eta^2_p = .22$. To investigate this interaction, we performed the same paired *t*-test for the song advantage as in the pitch deviation analyses. The contrast between speech and song was significant within each stimulus language condition ($p < .001$ in each case). This time, the song advantage was calculated by subtracting the average correlation in speech imitation from the average correlation in song imitation (the positive value reflects a song advantage) for each individual and each stimulus language. The song advantage for English stimuli (the yellow arrow in Figure 4) was significantly larger than that for Mandarin stimuli (the green arrow in Figure 4), *t*(126) = 5.67, *p* < .001, as was the case for pitch deviation scores. The same pattern of results was observed for both male and female participants.

The ANOVA also yielded a significant main effect of language background, *F*(1, 123) = 4.31, *p* = .040, η^2_p = .03. Overall, Mandarin speakers had higher pitch correlations (*M* = .76, *SD* = .08) than English speakers (*M* = .74, *SD* = .08). There was also a significant stimulus language × language background interaction, *F*(1, 123) = 5.14, *p* = .025, η^2_p = .04, and a significant stimulus domain × stimulus language × language background interaction, *F*(1, 123) = 8.82, *p* = .004, η^2_p = .07. There was no main effect of musical background and no other interactions in the pitch correlation analyses.

To interpret the three-way interaction, we performed two-way ANOVAs for each stimulus domain separately in order to focus on the role of language background for song and speech imitation. Figure 5 illustrates the means and standard deviations of each language background group for speech (left) and song (right) imitation. For speech imitation, there was a main effect of stimulus language, *F*(1, 123) = 75.54, *p* < .001, η^2_p = .38, and a significant language background × stimulus language interaction, *F*(1, 123) = 9.85, *p* = .002, η^2_p = .07. Planned comparisons between English and Mandarin speakers yielded a significant difference for Mandarin speech imitation, *t*(98.05) = -2.26, *p* = .026, but not for English speech imitation, *p* = .96. Mandarin speakers had higher pitch correlations for Mandarin speech (*M* = .72, *SD* = .11) than English speakers (*M* = .67, *SD* = .12) while this advantage was absent for English speech. On the other hand, for song imitation, there was a significant main effect of stimulus language, *F*(1, 123) = 7.02, *p* = .009, η^2_p = .05, language background, *F*(1, 123) = 5.98, *p* = .016. η^2_p = .05, but no interaction, *p* = .28. Pitch correlations for Mandarin song were slightly higher (*M* = .84, *SD* = .07) than those for English song (*M* = .83, *SD* =.08), regardless of participants' language background. Planned comparisons between English and Mandarin speakers also yielded a significant difference between two groups for both English song imitation, *t*(97.70) = -2.21, *p* = .029, and Mandarin song imitation, *t*(99.42) = -2.06, *p* = .041. Mandarin speakers had higher pitch correlations for songs of both stimulus languages (*M* = .84, *SD* = .07) than English speakers (*M* = .81, *SD* = .08).

Figure 5
Pitch correlation between the imitation and target pitch for speech imitation (left) and song imitation (right).



Note: Each dot indicates each participant's mean scores (blue = English speakers, red = Mandarin speakers, open = non-musicians, and closed = musicians). The bold lines show the group means and upper and lower lines represent 95% confidence intervals.

As we did for pitch deviation scores, we next addressed whether results for pitch correlations reflect variability that is independent of differences in perceptual ratings. Pitch correlations were significantly associated with perceptual ratings, based on a correlation modeled after the one we used for pitch deviations, *r*(11545) = .30, *p* < .001. The ANOVA on detrending scores again retained the critical main effect of language background, *F*(1,123) = 8.15, *p* = .005, η^2_p = .062, the stimulus domain × stimulus language interaction, *F*(1,123) = 17.11, *p* < .001, η^2_p

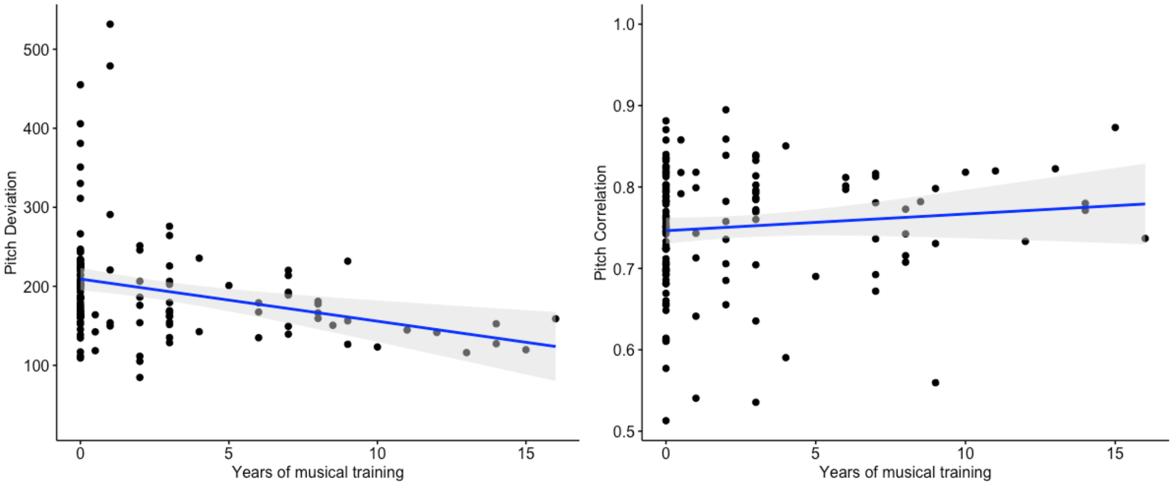
= .122, and the language background × stimulus language interaction, $F(1,123) = 8.25$, $p = .005$, $\eta^2_p = .063$. The only effect that no longer remained significant in the detrended data, again, was the main effect of stimulus domain (detrended $p = .791$, $\eta^2_p = .001$).

Correlation Analyses on Musical Background and pitch imitation

To further examine the effect of musical background on pitch imitation, we also performed Spearman's rank correlation analyses to estimate the relationship between years of musical training and pitch imitation accuracy. This exploratory analysis allowed us to examine the relationship between musical training and pitch imitation accuracy without dichotomizing musician vs non-musician groups based on a categorical cut-off (three years in the current study). Figure 6 shows the relationship between years of musical training and mean pitch deviation (left) and between years of musical training and mean pitch correlation (right). Both regression analyses were consistent with the results from the ANOVAs described above; years of musical training significantly correlated with mean pitch deviation, $r_s(125) = -.35$, $p < .001$, but not with mean pitch correlation, $p = .18$. Participants with more musical training matched the target pitch more accurately than those with less training.

Figure 6

The relationship between years of musical training and absolute pitch deviation (left) and between years of musical training and pitch correlation (right).



Note: Each dot represents a participant’s mean score. The filled areas represent 95% confidence intervals.

We also performed the same regression analyses for each stimulus categories (i.e., English speech, English song, Mandarin speech, and Mandarin song). Table 3 shows the Spearman’s correlation value and *p*-value for each stimulus category. We found similar trends as observed for the relationships shown in Figure 6 above: there was a significant association between years of musical training and mean pitch deviation for Mandarin speech and both types of song, but not English speech, whereas mean pitch correlation was not significantly correlated with years of musical training. These results further suggest that musical training is associated with an individual’s accuracy imitating absolute pitch but does not predict an individual’s accuracy imitating the pattern of relative pitch across the duration of music-like and speech-like stimuli.

Table 3

Spearman's correlation values and p values between years of musical training and mean pitch deviation/correlation for each stimulus category.

	Pitch Deviation			Pitch Correlation	
Stimulus Category	r_s	p		r_s	p
English Speech	-.16	.07		.11	.21
Mandarin Speech	-.25	<.01	**	.05	.63
English Song	-.32	<.001	***	.16	.08
Mandarin Song	-.38	<.001	***	.09	.33

Note: * $p < .05$, ** $p < .01$, *** $p < .001$

Discussion

The present study investigated whether the acoustic structures in speech and song, as well as differences in language and music background, influence pitch imitation ability. Two kinds of language stimuli (Mandarin and English) and two sets of linked song stimuli were included. Mandarin and English speakers with varying levels of musical training vocally imitated pitch sequences representative of speech and song. Results are interpreted in the context of stimulus based ('bottom-up') and participant background ('top-down') factors.

With respect to stimulus features, both measures of pitch imitation accuracy collectively demonstrated the song advantage over speech in pitch imitation for both native English and Mandarin speakers, which is consistent with the findings from previous studies (Mantell & Pfordresher, 2013; Wisniewski et al., 2013; F. Liu et al., 2013). This advantage might be affected by bottom-up factors (i.e., acoustic characteristics of song vs. speech). First, speech has greater pitch variability (Pfordresher et al., 2022; Stegmöller et al., 2008) and tends to be produced faster than music (Ding et al., 2017; Patel, 2014). Such structural differences might have conferred an advantage for song imitation – with less complex f_0 information and more time to encode and reproduce the target stimulus. Second, music often requires more perceptual precision in pitch processing (e.g., distinction between C and C#) than speech (Patel, 2014; Peretz & Hyde, 2003;

Zatorre & Baum, 2012), which in turn may help the vocalist detect and imitate pitch patterns in song better than those in speech.

It is important to consider the possibility that a song advantage could be also influenced by top-down factors (e.g., categorizations of speech vs song). In fact, we found a correlation between the perceptual rating of stimuli and the imitation performance. We addressed this issue in follow-up analyses that used linear detrending to address the degree to which vocal pitch imitation exhibits patterns that are distinct from what is revealed by perceptual ratings, using data from our perceptual rating task (see Table 2). Most results from the original analyses were preserved in the detrended data, suggesting that production effects are not reducible to processes involved in perceptual ratings. The only effect that exhibited redundancy across perception and production was the overall song advantage, which was no longer significant in detrended production data. However, it is important to note that the song advantage itself is subject to more nuance than is discernable by this present design. Recent analyses suggest that the song advantage can reverse depending on the timescale at which accuracy is assessed (Pfordresher, 2022). Thus, although the song advantage was redundant with perceptual ratings in the current study, it is doubtful that this redundancy would hold under all possible analyses. More important, the novel results from this study, reflecting the interplay between stimulus-driven (bottom-up) and experience-based (top-down) factors, exhibited effects in production that appear to be independent of perceptual responses.

We also observed more accurate imitation of relative pitch patterns (viz. pitch correlation measures) in songs than speech whereas Mantell and Pfordresher (2013) did not find this advantage. However, this null finding in Mantell and Pfordresher (2013) might be due to the smaller number of participants in that study (e.g., *N* = 25 in Experiment 1). Pfordresher (2022) combined and re-analyzed data from two previous studies (Mantell & Pfordresher, 2013; Pfordresher et al., 2021) and found a song advantage at a large time scale (i.e., across syllables and notes) using pitch correlations. Consistent with this result, the current study suggests the advantage of sung sequences for the relative pitch patterns.

Interestingly, the song advantage was larger for English stimuli than Mandarin stimuli, regardless of the language background of the participants. This suggests that pitch imitation accuracy can be influenced by the acoustic differences within a domain. Since the song stimuli in both languages consisted of the same structure (i.e., stable notes) and the accuracy for the song stimuli was similar across stimulus languages, here we focus on the differences in speech stimuli which might have determined the degree of the song advantage. Mandarin intonations consist of prototypical pitch movements (i.e., level, rising, falling-rising, and falling) that might be easier to imitate for even non-native speakers while English intonations do not have such prototypes. Because of their pitch structures, it is possible that Mandarin tones are perceived as more "music-like" than non-tonal languages especially by non-native speakers. In fact, our rating results (see *Materials*) showed that English speakers perceived Mandarin speech stimuli as more song-like than English speech stimuli. This perceptual difference might have resulted in more accurate imitation for Mandarin speech, similar to the song imitation. However, analyses of production that removed variability associated with perceptual ratings suggest that the advantage for Mandarin speech cannot be fully explained by recourse to perceptual ratings. Further studies are necessary to investigate the relationship between perception and imitation of various pitch structures.

We also addressed the role of top-down factors (i.e., musical and language background) on pitch imitation accuracy. The pitch deviation analyses showed that musical background influenced pitch imitation accuracy; participants who had at least three years of musical training matched the target pitch in both song and speech stimuli more accurately than those who did not. When evaluating musical training along a continuum, we found complementary results in that musical training was associated with pitch deviations scores for both speech (Mandarin stimuli) and song (Mandarin and English based stimuli). These results are in line with the OPERA hypothesis, which suggests that musical training facilitates speech processing by strengthening the neural networks involved in both music and speech domains (Patel, 2011). The current study further provides evidence for the benefits of musical training that transfers to the speech domain, specifically for absolute pitch matching ability. However, the pitch correlation analyses did not show the same robust benefit of musical training on pitch imitation accuracy. Perhaps, musical training facilitates a certain ability, such as pitch matching and pitch detection (Schön et al., 2004), but may not transfer reliably to ability that deviates from their initial training, such as imitation of relative pitch. In fact, the OPERA hypothesis specifically mentions that the benefits of musical training are based on the acoustic features that are focused on during musical training (Patel, 2011). In Western music, producing intended pitches (i.e., notes in diatonic scale) is a critical element of musical experience, especially in group musical activities. Musical training in such a way may facilitate individuals' ability to perceive and produce exact pitches, but this facilitation may not extend to the ability to produce relative pitch more accurately.

An alternate interpretation, brought up by an anonymous reviewer, is that musical training may encourage participants to perceive music-like qualities in speech, comparable to the effects of repetition in the speech to song illusion (Deutsch et al., 2011), particularly given recent results suggesting that hearing this illusion benefits vocal pitch imitation (Chen et al., in prep). Although trained musicians are not more prone to this illusion than untrained musicians, musically trained participants do tend to hear music-like qualities in speech overall (Vanden Bosch der Nederlanden et al., 2015), and that may be the case here. The specific basis for the musician advantage – whether based on neural connections for pitch processing or more high-level interpretations – will be addressed in forthcoming studies.

We were also interested in how language background influences the pitch imitation accuracy. The pitch correlation analyses revealed partial support for the effects of language background on imitation accuracy. First, Mandarin speakers imitated relative pitch patterns in song stimuli more accurately than English speakers, which is consistent with Pfordresher and Brown (2009) and Bidelman and colleagues (2011). Second, Mandarin speakers imitated the relative pitch patterns in Mandarin speech, but not English speech, more accurately than native English speakers. The reason why Mandarin background did not provide an advantage for English speech imitation might be explained by top-down factors that facilitate pitch processing. Since there was no phonetic information in the speech stimuli, perhaps the only cue participants had for better imitation was speech prosody. As mentioned in the beginning, Mandarin speech utilizes pitch patterns (i.e., lexical tones) to convey word meaning, whereas English speech does not have such tonal distinctions. Mandarin speakers have formed top-down knowledge of pitch patterns in their speech, which might have brought the advantage for Mandarin speech imitation. However, since there were no such lexical tones in English speech stimuli, the advantage of Mandarin background did not extend to the imitation of English speech. Finally, the effect of language background was not found in the pitch deviation analyses, which is consistent with Pfordresher and Brown (2009), who also found a tone language advantage only for the reproduction of relative pitch (in that study, relative pitch was measured on a note-by-note basis for melodies). One possible explanation for this result is that Mandarin language uses pitch movements to discriminate word meanings (i.e., contour-tone language) instead of level tones (i.e., register-tone language in which relative pitch height connotes lexical status, such as Cantonese). This practice, in turn, might have developed the ability to imitate the contours of target pitches, but not the exact pitches. Future studies should include native speakers of a register-tone language to examine the effect of their language background on absolute pitch matching.

One might suspect that bilingual/monolingual background, rather than tonal language background, might influence the vocal imitation ability. In the current study, all the Mandarin-speaking participants were bilingual/multilingual speakers who were fluent in English whereas about half of the English-speaking participants (50.98%) were bilingual or multilingual. However, when we analyzed data by comparing bilingual/multilingual (N = 26) and monolingual (N = 25) background in the English group, there was no effect of language background on the imitation accuracy, *p* = .83 for pitch deviation and *p* = .98 for pitch correlation. Future research should further address this issue by restricting the language background criteria, such as bilinguals of two non-tonal languages vs. bilinguals of a non-tonal language and a tonal language.

Although we present evidence that supports beneficial transfer of musical ability to speech imitation, we cannot eliminate the possibility of genetic factors on musical training. For example, Mosing and colleagues (2014) suggests genetic influence on the amount of music practice and musical ability (i.e., rhythm, melody, and pitch discrimination; Mosing et al., 2014). However, several longitudinal studies suggest causal effects of musical training on auditory processing (Habibi et al., 2018; Hyde et al., 2009). Hyde and colleagues (2009) demonstrated brain changes that correlated with improved motor and auditory skills after 15 months of musical training among children, suggesting structural brain differences between musicians and non-musicians might be due to musical training rather than genetic predispositions. Thus, these studies provide evidence for the effects of top-down factors on behavioral and neural differences. To further investigate the causational effects of musical training on pitch imitation ability, future studies would need to conduct a longitudinal experiment in which participants receive a musical training and pitch imitation tasks before and after the training.

To conclude, the current study addressed whether acoustic features as well as long-term experience in music and language affect the ability to vocally imitate pitch patterns in speech and songs. The main purpose of the current study was to investigate top-down and bottom-up factors involved in vocal production. Results suggest that pitch imitation accuracy is affected by the interaction of bottom-up and top-down factors. That is, different top-down factors influence vocal imitation ability for different dimensions of bottom-up factors (i.e., pitch structures): Musical background influences the ability to match absolute pitch whereas Mandarin language background influences the ability to match relative pitch patterns.

Acknowledgement

This work was funded by NSF grant BCS-1848930. FL was supported by a European Research Council (ERC) Starting Grant (CAASD, 678733). We thank Esther Song, Chantel Fatorma, Kaithlyn Massiah, David Vollweiler, Kayden Koh, Swathi Das, and Jonathan Jun Kit Liow for assistance with data processing and data collection and thank Li (Alice) Wang, Chen Zhao, Anna Gentile, and Kyle Walsh for assistance with stimuli creation.

References

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

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

Besson, M., Schön, D., Moreno, S., Santos, A, & Magne, C. (2007). Influence of musical expertise and musical training on pitch processing in music and language. *Restorative Neuology and Neuroscience, 25,* 399-410.

Bidelman, G. M., Hutka, S., & Moreno, S. (2013). Tone Language Speakers and Musicians Share Enhanced Perceptual and Cognitive Abilities for Musical Pitch: Evidence for Bidirectionality between the Domains of Language and Music. *PLoS ONE, 8,* e60676.

Bidelman, G. M., Gandour, J. T., & Krishnan, A. (2011). Musicians and tone-language speakers share enhanced brainstem encoding but not perceptual benefits for musical pitch. *Brain and Cognition, 77*(1), 1-10.

Boersma, P., & Weenink, D. (2013). *Praat: doing phonetics by computer* [Computer program]. http://www.praat.org/

Bregman, A. S. (1990). *Auditory Scene Analysis, The Perceptual Organization of Sound.* Cambridge, MA: MIT Press.

Deutsch, D., Henthorn, T., and Lapidis, R. (2011). Illusory transformation from speech to song. J*ournal of the Acoustical Society of America, 129,* 2245-2252.

Ding, N., Patel, A. D., Chen, L., Butler, H., Luo, C., & Poeppel, D. (2017). Temporal modulations in speech and music. *Neuroscience & Biobehavioral Reviews, 81,* 181-187.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39,* 179-191.

Giuliano, R., Pfordresher, P. Q., Stanley, E., Narayana, S., & Wicha, N. (2011). Native experience with a tone language enhances pitch discrimination and the speed of neural responses to pitch change. *Frontiers in Psychology, 2,* 146.

Kraus, N., & Chandrasekaran, B. (2010). Music training for the development of auditory skills. *Nature Reviews Neuroscience, 11*(8), 599–605.

Lee, C. Y., & Hung, T. H. (2008) Identification of Mandarin tones by English-speaking musicians and nonmusicians. *The Journal of the Acoustic Society of America, 124,* 3235–3248.

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

Liu, J., Hilton, C. B., Bergelson, E., & Mehr, S. A. (2023). Language experience predicts music processing in a half-million speakers of fifty-four languages. *Current Bioloty, 33,* 1-10.

Madsen, S. M. K., Marschall, M., Dau, T., & Oxenham, A. J. (2019). speech perception is similar for musicians and non-musicians across a wide range of conditions. *Science Reports, 9*(1), 10404.

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

Mercado, E., III, Mantell, J. T., & Pfordresher, P. Q. (2014). Imitating sounds: A cognitive approach to understanding vocal imitation. *Comparative Cognition & Behavior Reviews, 9,* 1-57.

Mosing, M. A., Madison, G., Pedersen, N. L., Kuja-Halkola, R., & Ullén, F. (2014). Practice Does Not Make Perfect: No Causal Effect of Music Practice on Music Ability. *Psychological Science, 25,* 1795-1803.

Nahum, M., Nelken, I., & Ahissar, M. (2008). Low-Level Information and High-Level Perception: The Case of Speech in Noise. *PLoS Biology, 6*, e126.

Ong, J., Wong, P. C. M., & Liu, F. (2020). Musicians show enhanced perception, but not production, of native lexical tones. *The Journal of the Acoustical Society of America, 148*(6), 3443-3454.

Ozaki, Y., Tierney, A., Pfordresher, P. Q., McBride, J., Benetos, E., Proutskouva, P., Chiba, G., Liu, F., Jacoby, N., Purdy, S. C., Opondo, P., Fitch, W. T., Rocamora, M., Thorne, R., Nweke, F., Sadaphal, D., Sadaphal, P., Hadavi, S., Fujii, S., … Savage, P. E. (accepted in principle). Globally, songs are slower, higher, and use more stable pitches than speech [Stage 2 Registered Report]. *Peer Community in Registered Reports. Preprint:* https://doi.org/10.31234/osf.io/jr9x7

Patel, A. D. (1998). Syntactic processing in language and music: Different cognitive operations, similar neural resources? *Music Perception, 16*(1), 27–42.

Patel, A. D. (2011) Why would musical training benefit the neural encoding of speech? The OPERA hypothesis. *Frontier, 2*(142), 1-14.

Patel, A. D. (2014). Can nonlinguistic musical training change the way the brain processes speech? The expanded OPERA hypothesis. *Hearing Research, 308*, 98-108.

Peretz, I., & Coltheart, M. (2003). Modularity of music processing. *Nature Neuroscience, 6,* 688-691.

Peretz, I., & Hyde, K. (2003). What is specific to music processing? Insights from congenital amusia. *Trends in Cognitive Sciences, 7*(8), 362-367.

Pfordresher, P. Q. (2022). A reversal of the song advantage in vocal pitch imitation. *JASA Express Letters, 2*(3), 034401

Pfordresher, P. Q., & Brown, S. (2009). Enhanced production and perception of musical pitch in tone language speakers. *Attention Perception & Psychophysics, 71,* 1385-1398.

Pfordresher, P. Q., & Chow, K. (2019). A cost of musical training? Sensorimotor flexibility in musical sequence learning. *Psychonomic Bulletin & Review, 26,* 967-973.

Pfordresher, P. Q., & Demorest, S. M. (2021). The prevalence and correlates of accurate singing. *Journal of Research in Music Education, 69,* 5-23.

Pfordresher, P. Q., Mantell, J. T., & Pruitt, T. A. (2022). Effects of intention in the imitation of sung and spoken pitch. *Psychological Research, 86,* 792-807.

Schön, D., Magne, C., & Besson, M. (2004) The music of speech: Music training facilitates pitch processing in both music and language. *Psychophysiology, 41,* 341-349.

Stegemöller, E., Skoe, E., Nicol, T., Warrier, C. M., & Kraus, N. (2008). Musical training and vocal production of speech and song. *Music Perception, 25,* 419-428.

Tan, Y. T., McPherson, G. E., Peretz, I., Berkovic, S. F., & Wilson, S. J. (2014). The genetic basis of music ability. *Frontiers in Psychology, 5,* 658.

Thompson WF, Schellenberg EG, Husain G (2004) Decoding speech prosody: Do music lessons help? *Emotion, 4,* 46–64.

Van de Cavey, J., & Hartsuiker R. J. (2016). Is there a domain-general cognitive structuring system? Evidence from structural priming across music, math, action descriptions, and language. *Cognition, 146,* 172-184.

Vanden Bosch der Nederlanden, C. M., Hannon, E. E., & Snyder, J. S. (2015). Everyday musical experience is sufficient to perceive the speech-to-song illusion. *Journal of Experimental Psychology: General, 144*, e43-e49.

Wang, Y., Jongman, A., & Sereno, J. A. (2001). Dichotic perception of Mandarin tones by Chinese and American listeners. *Brain and Language, 78,* 332–348.

Wang, Y., Sereno, J. A., Jongman, A., & Hirsch, J. (2003). fMRI Evidence for Cortical Modification during Learning of Mandarin Lexical Tone. *Journal of Cognitive Neuroscience, 15*(7), 1019-1027.

Wang, L., Pfordresher, P. Q., Jiang, C., & 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 Research, 14,* 2355-2372.

Wisniewski, M. G., Mantell, J. T., & Pfordresher, P. Q. (2013). Transfer effects in the vocal imitation of speech and song. *Psychomusicology: Music, Mind and Brain, 23,* 82-99.

Wong, P. C. M., Skoe, E., Russo, N. M., Dees, T., & Kraus, N. (2007). Musical experience shapes human brainstem encoding of linguistic pitch patterns. *Nature Neuroscience, 10,* 420-422.

Yip, M. (2002). *Tone.* Cambridge: Cambridge University Press.

Zatorre, R. J., & Baum, S. R. (2012). Musical Melody and Speech Intonation: Singing a Different Tune. *PLoS Biol, 10,* e1001372.

Zatorre, R. J., & Gandour, J. T. (2008). Neural specializations for speech and pitch: moving beyond the dichotomies. *Philosophical Transactions of the Royal Society B: Biological Sciences, 363,* 1087-1104.

Zatorre, R. J., Belin, P., & Penhune, V. B. (2002). Structure and function of auditory cortex: Music and speech. *Trends in Cognitive Sciences, 6,* 37-46.

Appendix A

The list of speech stimuli

<u>English Sentences</u>

1. They **like** to **bike** home.
   They **like** to **bike home?**
2. I **saw** a **new** car.
   I **saw** a **new car?**
3. The **children** can’t **sleep.**
   The **children** can’t **sleep?**
4. The **dog** ate.
   The **dog ate?**
5. The **door** is **blue.**
   The **door** is **blue?**
6. The **boys** play **golf.**
   The **boys** play **golf?**
7. **Jane** went **back.**
   Jane went **back?**
8. **Nick** ran away.
   Nick ran **away?**
9. The **shirt** smells.
   The **shirt smells?**
10. The **trees** are green.
    The **trees** are **green?**
11. **We** drink **water.**
    We drink **water?**
12. **Wheels** are **round.**
    Wheels are **round?**

<u>Mandarin Sentences</u>

1. **您**几点睡觉。
   您几点**睡觉**？
2. **梅**挽回十分。
   梅挽回**十分**？
3. 这夏**天**特闷。
   这夏**天**特**闷**？
4. 猫不**多。**
   猫**不多？**
5. 他一米**八**。
   他一米**八？**
6. 你**来**我这儿。
   你**来**我**这儿？**
7. **梅**如画。
   梅如**画？**
8. **猫**爱吃冰。
   猫爱吃**冰**？
9. 黄**花**开。
   黄花**开**？
10. 让**他**别唱。
    让他别**唱？**
11. **他**让**您**来。
    他**让您来**？
12. **没**来过。
    没来**过**？

Note: The meaning of each English sentence (left) does not correspond to that of the Mandarin sentence (right). The speakers were instructed to emphasize the syllables written in bold.

Appendix B

Acoustic features of stimuli

Analyses of acoustic features focused on properties of pitch rather than timing. All stimuli were matched with respect to duration, and the removal of phonetic information renders ambiguous the perceptual location of note or syllable onsets. Features associated with timing – tempo (related to overall duration) and rhythm (relative timing of event onsets) – thus were not considered to be informative.

Table B.1. Displays means and standard deviations for two key acoustic features: The mean pitch height of each item (M f_0), and variability of pitch (SD f_0), with both measures computed across the entire sequence. Potential differences across the two critical stimulus factors (Stimulus Language and Stimulus Domain) were analyzed using ANOVAs for each feature, using item as the random factor (N = 48 for each cell) and treating the stimulus factors as between-"subjects" variables.

Table B.1
Means (standard deviations) acoustic features by stimulus Language and Stimulus Domain

		Target Statistics			
Target Category		Pitch height		Pitch variability	
Language	Domain	*M*	*SD*	*M*	*SD*
English	Speech	190.20	47.82	42.74	18.28
English	Song	197.34	67.72	39.24	26.71
Mandarin	Speech	195.83	58.04	36.99	15.93
Mandarin	Song	199.71	57.47	38.83	18.96

No main effects or interactions were significant for either dependent variable (all $p > .25$, all $\eta^2_p < .005$). This may be surprising in light of other results showing higher pitch height for song and speech and more unstable pitch for speech than song (e.g,. Ozaki et al., accepted in principle). However, these null effects do follow from two important considerations. First, the manner in which pitches were mapped to syllables in the present stimuli controlled for any overall differences in pitch height. Second, the fact that pitch variability here was aggregated across the entire sequence (which is necessary based on the stimulus construction) resulted in variability being attributable to both variability on a small timescale (characteristic of speech e.g., within syllables) and larger timescales (e.g., change across notes in song).

Appendix C

Analysis addressing unequal N across groups

The sampling strategy we used (see *Participants*), led to unequal numbers of participants in the four groups, as shown in Table 1. This deviation from the assumptions of ANOVA leaves the analyses we report open to the possibility that significant effects from our grouping variables (Language Background, Musical Background) are vulnerable to Type I errors. We addressed this problem through a permutation test based on Pfordresher and Chow (2019). On each of 1,000 permutations, group labels were randomly shuffled across participants. Specifically, 83 participants were randomly labeled ‘non-musician’ and the remaining 44 were labeled ‘musician’. Likewise, 51 participants were randomly labeled ‘English speakers’ and the remaining 76 were labeled ‘Mandarin speakers’. An ANOVA using the same design reported in the primary study was then conducted, using these randomly chosen designations for the grouping variables. If the unequal sample sizes across groups lead to false positives, then we should find significant effects associated with these randomly chosen groups on more than 5% of permutations.

For each dependent variable, we focused on significant effects associated with group variables in the primary study. The main effect of musical training on pitch deviation scores, significant in the primary study, was only significant on 48 permutations with randomly assigned group labels (p = .048). The significant musical training × stimulus domain interaction, also significant in the primary study, was only significant on 42 permutations with randomly assigned group labels (p = .042). We concluded that these effects of musical background on pitch deviations in the primary study were unlikely to have arisen as a byproduct of unequal sample sizes.

The main effects of language background on pitch correlation scores, significant in the primary study, was only significant on 45 permutations with randomly assigned group labels (*p* = .045). Likewise, the significant language background × stimulus domain interaction was only significant for 45 permutations with randomly assigned labels (*p* = .045), and the language background × stimulus domain × stimulus language interaction was only significant for 46 permutations with randomly assigned labels (*p* = .046). We concluded that these effects of language background on pitch correlations in the primary study were unlikely to have arisen as a byproduct of unequal sample sizes because the number of permutations that were significant did not exceed 5% in any case.