Elucidating the Influences of Embodiment and Conceptual Metaphor on Lexical and Non-Speech Tone Learning

Laura M. Morett
Jacob B. Feiler
Dept. of Educational Studies in Psychology, Research Methodology, and Counseling
University of Alabama

Laura M. Getz
Dept. of Psychological Sciences
University of San Diego
Author Note

This work was supported by a *Language Learning* Early Career Award and a Level 1 Research Grant from the University of Alabama Research Grants Council to L.M.M. and a Graduate Council Fellowship and Dean’s Summer Research Fellowship to J.B.F. All data and analysis scripts are publicly available via the Open Science Framework at [https://osf.io/2wd87/](https://osf.io/2wd87/).

The authors thank Yi Chen and Jin Zhao for assistance with recording of audio and video stimuli and Katherine Brockmann, Raymond Bulger, and Mary Bray for assistance with paradigm programming and data collection. In addition, the authors thank the attendees of the Southeastern Psychological Association, Southeastern Universities Graduate Symposium, and University of Alabama Research and Creative Activity Conference and Faculty Research Day for their feedback on portions of this research.

Correspondence concerning this article should be addressed to Laura M. Morett, Box 870240, University of Alabama, Tuscaloosa, AL 35401. Email: lmorett@ua.edu
Abstract

In the contexts of language learning and music processing, hand gestures conveying acoustic information visually influence perception of speech and non-speech sounds (Connell et al., 2013; Morett & Chang, 2015). Currently, it is unclear whether this effect is due to these gestures’ use of the human body to highlight relevant features of language (embodiment) or the cross-modal mapping between the visual motion trajectories of these gestures and corresponding auditory features (conceptual metaphor). To address this question, we examined identification of the pitch contours of lexical tones and non-speech analogs learned with pitch gesture, comparable dot motion, or no motion. Critically, pitch gesture and dot motion were either congruent or incongruent with the vertical conceptual metaphor of pitch. Consistent with our hypotheses, we found that identification accuracy increased for tones learned with congruent pitch gesture and dot motion, whereas it remained stable or decreased for tones learned with incongruent pitch gesture and dot motion. These findings provide the first evidence that both embodied and non-embodied visual stimuli congruent with the vertical conceptual metaphor of pitch enhance lexical and non-speech tone learning. Thus, they illuminate the influences of conceptual metaphor and embodiment on lexical and non-speech auditory perception, providing insight into how they can be leveraged to enhance language learning and music processing.

Keywords: Multisensory integration; conceptual metaphor; embodied cognition; lexical tone; gesture
Elucidating the Influences of Embodiment and Conceptual Metaphor on Lexical and Non-Speech Tone Learning

Learning novel speech sounds is a challenging aspect of second language (L2) acquisition, especially for adult learners (Flege, 1995; Tyler et al., 2014). A parameter of speech sounds that poses particular difficulties for English speakers is lexical tone (Pelzl, 2019), which consists of pitches distinguishing between word meanings or inflections (Gussenhoven, 2004; Maddieson, 2013; Yip, 2002). In Mandarin, the most widely-spoken tonal language, there are four lexical tones characterized by the heights and contours of their fundamental frequency (F0): (1) high-flat; (2) rising; (3) low or low-dipping; (4) falling (Chao, 1965; Ho, 1976; Howie, 1974). For native Mandarin speakers, detecting differences between lexical tones primarily entails discriminating between pitch contours. By contrast, native speakers of atonal languages, such as English, tend to rely on pitch height to differentiate between Mandarin lexical tones (Francis et al., 2008; Gandour, 1983; Huang & Johnson, 2010). Thus, promoting successful acquisition of Mandarin lexical tones in speakers of English and other atonal languages involves conveying the importance of the differences between their pitch contours. Learning to discriminate between Mandarin lexical tones based on their pitch contours may contribute to word recognition in native atonal language speakers (Morett & Chang, 2015; Wong & Perrachione, 2007), although recent work suggests that even advanced atonal L2 Mandarin learners may find it challenging to discriminate between Mandarin words and nonwords based solely on lexical tone (Han & Tsukada, 2020; Ling & Grüter, 2020; Pelzl et al., 2019, 2021).

Previous research has shown that pitch gestures—metaphorical gestures consisting of upwards and downwards hand movements reflecting high- and low-frequency pitch movements, respectively—enhance English speakers’ differentiation between Mandarin lexical tones by
conveying their pitch contours visually and haptically via motion (Baills et al., 2019; Morett & Chang, 2015; Zhen et al., 2019; Zheng et al., 2018). At present, however, it is unclear whether this effect is due to these gestures’ use of the human body to convey the pitch contours of lexical tones haptically (embodiment) or the cross-modal mapping between the visual motion trajectories of these gestures and corresponding auditory pitch contours (conceptual metaphor). Although conceptual metaphor is inherently grounded in embodied experiences (e.g., high pitch as high in space; Lakoff & Johnson, 1980), it is important to determine whether non-embodied stimuli conveying pitch contours (i.e., visual depictions) are as effective as their embodied counterparts (i.e., pitch gestures) as pedagogical aids to facilitate lexical tone learning. Moreover, it is unknown whether conceptual metaphor and embodiment affect acquisition of non-speech sounds (e.g., musical tones) similarly to lexical tone. Comparing how embodied and non-embodied visual depictions of pitch contours affect acquisition of speech and non-speech tones will provide insight into their ability to enhance learning and perception of musical tones in addition to lexical tones, broadening the scope of their application beyond L2 acquisition. The current study addresses both of these outstanding questions, elucidating the impacts of embodiment and conceptual metaphor on lexical and non-speech tone learning. In what follows, we provide an overview of how conceptual metaphor and embodiment influence lexical tone learning, then discuss similarities and differences between speech and non-speech tones, and finally review evidence for overlap between lexical and musical tone perception before detailing the goals, hypotheses, and contributions of the current study.

**Conceptual Metaphor and Embodiment in Lexical Tone Learning**

Despite the challenges that many adult native atonal language speakers encounter in perceiving and differentiating between lexical tones, these tones can nevertheless be learned
successfully with sufficient training (Wang et al., 1999). Some research indicates that learning Mandarin lexical tones with visual depictions of their pitch contours enhances discrimination between them (Bluhme & Burr, 1971; Godfroid et al., 2017; Liu et al., 2011). Other research indicates that learning Mandarin words differing minimally (i.e., only) in lexical tone while observing pitch gestures enhances differentiation between them and, thus, their meanings (Baills et al., 2019; Morett & Chang, 2015; Zhen et al., 2019). Because the effects of visual depictions and gestures conveying pitch contours have not been directly compared, however, it is unclear at present whether one is more effective in facilitating L2 lexical tone learning than the other.

One theory that has been leveraged to explain why visual depictions and gestures conveying pitch contours enhance learning of the distinctions between Mandarin lexical tones is conceptual metaphor theory (Lakoff & Johnson, 1980). This theory postulates that abstract ideas are grounded in concrete embodied experiences, which are implicitly mentally simulated during comprehension and production. Conceptual metaphor theory has been applied to the domain of music to explain that pitch is based on a vertical conceptual metaphor in which the “up” direction represents high frequency pitch and the “down” direction represents low frequency pitch (Casasanto et al., 2003; Connell et al., 2013). Thus, it explains how auditory pitch frequency is mapped cross-modally onto position and motion in visual space. The vertical conceptual metaphor of pitch has also been applied to intonation (Bolinger, 1983; Kelly et al., 2017) and lexical tone (Baills et al., 2019; Morett & Chang, 2015; Zheng et al., 2018) in speech, particularly with respect to their acquisition in a novel L2. Importantly, although this metaphor can be mapped onto other planes (e.g., horizontal; Zhen et al., 2019), congruency with it is crucial; thus, we hypothesize that pitch gestures depicting contours of lexical tones other than those they occur with should not facilitate—and may in fact hinder—lexical tone learning.
Another theory that has been leveraged to explain why pitch gestures enhance learning of the distinctions between Mandarin lexical tones is embodied cognition, or embodiment (Shapiro, 2019). With respect to gesture, this theory postulates that the body is capable of conveying thought via meaningful action. Like conceptual metaphor theory, embodiment maintains that production and comprehension of gesture entails implicit mental simulation of embodied experiences with referents (Hostetter & Alibali, 2008; McNeill, 2005). This mental simulation may rely on the mirror neuron system, which is similarly active during action execution and observation (Rizzolatti & Craighero, 2004), particularly during gesture comprehension (Kelly et al., 2008), although there is disagreement concerning its stature as the underlying neural mechanism (Hickok, 2014). As applied to L2 lexical tone learning, embodiment maintains that pitch gestures utilize meaningful body action to convey the pitch contours of lexical tones haptically as well as visually, thereby facilitating their acquisition. The beneficial effect of pitch gesture on lexical tone acquisition demonstrates that embodiment and conceptual metaphor are not mutually exclusive and can in fact work in tandem. Indeed, embodiment based on a conceptual metaphor relevant to unfamiliar speech sounds may more effectively promote differentiation between words containing them than other forms of embodiment (e.g., gestures conveying the meanings of such words; Kelly & Lee, 2012; Morett & Chang, 2015).

Evidence that observing both images and gestures conveying the pitch contours of Mandarin lexical tones facilitates differentiation between them in adult English speakers (Baills et al., 2019; Bluhme & Burr, 1971; Godfroid et al., 2017; Liu et al., 2011; Morett & Chang, 2015; Zhen et al., 2019) suggests that conceptual metaphor may be more fundamental than embodiment via gesture for promoting discrimination between lexical tones. Therefore, non-
embodied dynamic visual stimuli based on the vertical conceptual metaphor of pitch (e.g., moving dots) may be just as effective as pitch gestures in this context.

**Lexical Versus Non-Speech Tones**

As mentioned previously, prior research provides evidence that conceptual metaphor and embodiment play a key role in pitch perception in non-speech (i.e., musical) tone, as well. With respect to conceptual metaphor, non-embodied visual depictions based on the vertical metaphor of pitch influence perception of pitch in non-speech sounds (Ben-Artzi & Marks, 1995; Casasanto et al., 2003; Evans & Treisman, 2010; Getz & Kubovy, 2018; Melara & O’Brien, 1987; Mossbridge et al., 2011; Patching & Quinlan, 2002), in addition to such depictions of other relevant conceptual metaphors of pitch (e.g., size, brightness; see Spence, 2011, for review).

With respect to embodiment, ascending and descending pitch are associated with upwards and downwards gesture production, respectively (Godøy et al., 2006; Küssner et al., 2014). Moreover, observation of vertical gestures influences pitch judgment for vocal musical notes, with downwards gestures leading to perception of pitch as lower and upwards gestures leading to perception of pitch as higher (Connell et al., 2013). However, no research to date has compared the effects of non-embodied visual pitch depictions and pitch gestures on discrimination training for lexical and non-speech tones, so it is currently unclear whether conceptual metaphor and embodiment via gesture affect their learning similarly.

More generally, research has investigated the extent to which non-speech and lexical tones are processed similarly, as processing of both relies on pitch perception. Indeed, both musical and lexical tones comprise multiple pitches (Tang et al., 2016) conveying information in a time-varying pattern (Chandrasekaran et al., 2009). Additionally, musical and lexical tones are generative, such that several small components (e.g., pitches and segments) combine to create a
complex sound interpreted as speech or song (Alexander et al., 2005). Like lexical tones, musical tones can also be distinguished based on small, distinct intervals (Tang et al., 2016). However, there are also abundant differences in the spectro-temporal properties (i.e., signal strength over time) of music and speech. For example, musical notes are characteristically longer in duration and have more variability than speech syllables (Ding et al., 2017). The change in spectral shape within musical notes is also much less drastic than the spectral change found within spoken syllables (Patel, 2011). These similarities and differences between lexical and musical tones have led to different theories of their processing, with more evidence favoring similar than distinct processing, as detailed below (see Bradley, 2013, for a review).

Liberman and colleagues (1967) proposed that because humans have an innate ability to perceive the sounds of speech, its structure is special. Phonology encodes an infinite number of words and sentences using as few units of speech as possible, allowing humans to integrate suprasegmentals, phonemes, and other temporal information into specific phonetic segments (see Liberman, 1984). Research supporting the speech is special hypothesis opposes the notion that perceiving speech is comparable to perceiving non-speech sounds (Liberman & Mattingly, 1989); instead, it suggests that speech perception involves interpretation of articulatory events rather than auditory and acoustic events. Specifically, perceiving speech involves implicitly mentally simulating movements of the vocal tract, tongue, lips, and jaw used to produce it known as articulatory gestures (Liberman & Mattingly, 1985). Associations between articulatory gestures and speech sounds are implicitly acquired via both listening and speaking, negating the need for intentional decoding and processing (Liberman et al., 1967).

Early support for the speech is special hypothesis rested on the assumptions of categorical perception, which posits that distinct phoneme boundaries exist in language despite
continuous input. Thus, acoustic differences within these boundaries are not discriminated as accurately as equivalent differences spanning them (Liberman et al., 1957). Contrary to the claims of this hypothesis, however, non-speech sounds can be perceived categorically (Burns & Ward, 1978; Xu et al., 2006), and non-human animals such as chinchillas, macaques, and quail perceive speech sounds categorically (Kluender et al., 1987; Kuhl & Miller, 1975; Kuhl & Padden, 1983). Moreover, humans are sensitive to within-category perceptual differences in speech (McMurray et al., 2002; Toscano et al., 2010), challenging the core claim of categorical perception.

Thus, domain-general models of speech perception such as general auditory models and cue-integration models posit that speech is perceived via continuous auditory contrasts (e.g., Diehl & Kluender, 1989; Massaro & Oden, 1980) or acoustic cues estimated from domain-general statistics (e.g., Getz et al., 2017; Toscano & McMurray, 2010), maintaining that speech is not special compared to other auditory inputs (Diehl et al., 2004; Samuel, 2011). Music perception research also supports this conclusion by demonstrating that music and speech are processed similarly (Juslin & Laukka, 2003; Koelsch, 2011; McMullen & Saffran, 2004). Specifically, the “O” in Patel’s (2011; 2014) OPERA hypothesis posits that because music and speech share both acoustic (e.g., pitch, timing, timbre) and cognitive (i.e., memory and attentional resource) properties, they are processed using overlapping cognitive and neural mechanisms.

The current study contributes to the debate over whether speech is processed via specialized or domain-general mechanisms by providing insight into the extent to which the effects of pitch gestures and non-embodied visual pitch depictions generalize to non-speech tone learning. If these effects are absent or significantly diminished or differ in direction for non-
speech tones, the results would lend credence to the speech is special hypothesis by suggesting that pitch is interpreted differently in music than in speech. On the other hand, if these effects are similar or amplified and are the same in direction for non-speech tones, the results would support domain-general hypotheses of speech perception by suggesting that pitch is interpreted similarly in music and speech.

**Evidence for Overlap Between Lexical and Musical Tone Processing**

Evidence for overlap in processing between speech and music comes from two different lines of work: research examining the impact of musical experience on speech perception and research comparing tonal and atonal language speakers on musical pitch perception.

In regard to the impact of musical experience on speech perception, musicians are more sensitive to pitch differences in speech than non-musicians (Chandrasekaran et al., 2009; Kraus & Chandrasekaran, 2010; Magne et al., 2006). For example, weak incongruities in pitch at the end of spoken phrases elicit event-related potentials more positive in amplitude in musicians than non-musicians (Besson et al., 2007; Marques et al., 2007). Moreover, musicians can successfully distinguish between Mandarin lexical tones even without prior experience with Mandarin (Chandrasekaran et al., 2009; Wong et al., 2007). These findings indicate that musical experience positively correlates with sensitivity to pitch in speech.

In regard to musical pitch perception in tonal versus atonal language speakers, relative to English speakers, Mandarin speakers show greater neural sensitivity to differences between non-speech sounds with pitch contours spanning Mandarin lexical tone categories and less neural sensitivity to differences between non-speech stimuli with pitch contours within Mandarin lexical tone categories (Chandrasekaran et al., 2007, 2009). Similarly, when identifying the pitch contours (e.g., rising, falling, flat) of non-speech lexical tone analogs, native Mandarin speakers
are more likely to misclassify flat and falling contours than native English speakers (Bent et al., 2006), suggesting that linguistic experience extends to non-speech processing under certain matching conditions. These findings are specific to non-speech sounds with pitch contours similar to Mandarin lexical tones, however. By contrast, Mandarin-speaking adults and children bilingual in Mandarin and an atonal language discriminate between non-speech pure tones with accuracy comparable to English-speaking adults and children bilingual in English and another atonal language, regardless of the pitches of these tones (Bent et al., 2006; Morett, 2020). Together, these findings indicate that long-term exposure to a tonal language affects perception of non-speech sounds similar in pitch to lexical tones in that language.

Taken together, the results of both of these lines of research provide evidence of substantial overlap between the cognitive and neural mechanisms used to process pitch in music and language, consistent with domain-general theories of speech perception. Thus, Mandarin lexical tones and non-speech stimuli with analogous pitch contours (i.e., analogs) are likely learned similarly by atonal language speakers, and conceptual metaphor and embodiment likely affect their learning similarly, as well.

**Current Study**

In the current study, we investigated the effects of conceptual metaphor and embodiment on the learning of lexical tones and non-speech analogs. To do so, we examined identification accuracy for these auditory stimuli before and after they were learned by observing pitch gesture, comparable dot motion, and no motion. In light of evidence that observing visual stimuli based on the vertical conceptual metaphor of pitch affects lexical tone acquisition (Baills et al., 2019; Bluhme & Burr, 1971; Godfroid et al., 2017; Liu et al., 2011; Morett & Chang, 2015; Zhen et al., 2019) as well as pitch perception in musical stimuli (Ben-Artzi & Marks, 1995; Casasanto et al.,
2003; Connell et al., 2013; Evans & Treisman, 2010; Getz & Kubovy, 2018; Godøy et al., 2006; Küßner et al., 2014; Melara & O’Brien, 1987; Mossbridge et al., 2011; Patching & Quinlan, 2002), we predicted that observing pitch gesture and dot motion at learning would elicit greater changes in tone identification accuracy from pre-test to post-test than observing no motion at learning. Moreover, based on the same evidence, we predicted that pre-test to post-test changes in identification accuracy would not differ significantly for tones learned with pitch gesture and dot motion. Because pitch gesture is embodied and therefore involves haptic as well as visual processing, observing it may result in more multimodal and therefore more robust mental representations than observing non-embodied dot motion, leading to stronger effects on tone learning. Thus, an alternative possibility is that, relative to observing dot motion, observing pitch gesture may elicit greater pre-test to post-test changes in tone identification accuracy.

Another question of interest was how congruency with the vertical conceptual metaphor of pitch would affect lexical tones and non-speech analog learning. Based on previous research demonstrating that non-embodied visual depictions as well as gestures consistent with the vertical conceptual metaphor of pitch enhance lexical tone acquisition (Baills et al., 2019; Bluhme & Burr, 1971; Godfroid et al., 2017; Liu et al., 2011; Morett & Chang, 2015; Zhen et al., 2019; Zheng et al., 2018), we predicted that observing pitch gesture and dot motion would influence tone learning in accordance with this conceptual metaphor. Specifically, we predicted that learning tones while observing pitch gesture and dot motion congruent with the vertical conceptual metaphor of pitch would increase the accuracy of their identification from pre-test to post-test by enabling cross-modal mapping of visual information concerning pitch onto auditory percepts of it, thereby strengthening mental representations of tones. Moreover, we predicted that learning tones while observing pitch gesture and dot motion incongruent with this vertical
conceptual metaphor would decrease the accuracy of their identification from pre-test to post-test by preventing this cross-modal mapping, thereby weakening mental representations of tones.

A final question of interest was whether conceptual metaphor and embodiment affect lexical tone and non-speech analog learning similarly. In light of evidence that the vertical conceptual metaphor of pitch and embodiment of it via gesture affects pitch perception in musical stimuli (Ben-Artzi & Marks, 1995; Casasanto et al., 2003; Connell et al., 2013; Evans & Treisman, 2010; Getz & Kubovy, 2018; Godøy et al., 2006; Küssner et al., 2014; Melara & O’Brien, 1987; Mossbridge et al., 2011; Patching & Quinlan, 2002) similarly to lexical tone acquisition (Baills et al., 2019; Bluhme & Burr, 1971; Godfroid et al., 2017; Liu et al., 2011; Morett & Chang, 2015; Zhen et al., 2019; Zheng et al., 2018), we predicted that the change in tone identification accuracy from pre-test to post-test would not differ significantly or would be amplified and would be the same in direction for analogs relative to lexical tones learned with visual stimuli based on the vertical conceptual metaphor of pitch, consistent with domain-general theories of speech perception (Diehl et al., 2004; Samuel, 2011). Alternatively, if these visual stimuli affect processing of non-speech analogs differently than lexical tones, their effects should be diminished significantly or should differ in direction for non-speech analogs relative to lexical tones, consistent with the speech is special hypothesis (Liberman & Mattingly, 1989).

**Method**

**Participants**

Sixty-four individuals between 18 and 35 years old (M = 20.45 S.D. = 3.66) from universities in the Southeastern and Western U.S. participated in this experiment in return for partial course credit. Participants were monolingual English speakers with no knowledge of Mandarin or any other tonal language. All participants reported normal or corrected-to-normal
vision and normal hearing. Participation in the experiment required completing one self-paced online experimental session.

**Materials**

Two lists of monosyllabic consonant-vowel(-consonant) (CV(C)) Mandarin word triplets consisting of phonemes present in English (see Appendix A) were compiled for use in the experiment. Seven triplets were included in each list to cover a diverse range of vowels, upon which lexical tone is superimposed, and initial consonants. Triplets were matched on vowels across lists, whereas initial consonants of triplets were not repeated within or across lists. Mandarin words within each triplet differed from one another only in lexical tone (e.g., ba1, ba2, ba4). As indicated in the previous section, there are four lexical tones in Mandarin, but tones 1, 2, and 4 have the clearest visual analogs (flat, rising, falling); thus, tone 3 was not included in the experiment (as in Bent et al., 2006; Chen et al., 2017).

All audio and video stimuli used in this study are publicly available via the following link: https://osf.io/2wd87/. Audio recordings of Mandarin words were made by recording a male native Mandarin speaker who was bilingual in English pronouncing them. After all the Mandarin words were recorded, they were segmented into individual audio files using Audacity (Version 2.3.0). To eliminate any differences in volume across individual recordings, the intensity of these audio files was normalized to 71.44 dB using Praat (Version 6.1.06; Boersma & Weenink, 2016). To ensure that Mandarin word stimuli were naturalistic, their durations were not altered and therefore differed across lexical tone categories.

Sine wave analogs corresponding to Mandarin words were generated with an online sweep tone generator (https://www.audiocheck.net/audiofrequencysignalgenerator_sweep.php) using a linear sine wave tone change. All analogs had onsets and offsets 0.05 s in length and
were standardized to an intensity of 70 dB. For the seven triplets in List 1, the starting pitches of tone 1 analogs varied between 200 Hz and 260 Hz in 10 Hz increments, based on the starting pitches of tone 1 Mandarin words. Likewise, for the seven triplets in List 2, the starting pitches of tone 1 analogs varied between 205 Hz and 265 Hz in 10 Hz increments, based on the starting pitches of tone 1 Mandarin words. For all 14 triplets, the percentage pitch changes used for tone 2 and tone 4 analogs were equivalent to the percentage pitch changes of tone 2 and tone 4 Mandarin words, respectively. Durations of analogs were based on durations of Mandarin words.

Table 1 contains means and standard deviations for durations of Mandarin words and analogs. To determine whether durations of auditory stimuli differed by tone and stimulus type, we submitted durations to a linear mixed effect model with fixed effects of tone and stimulus type and a random effect of auditory stimulus. This analysis revealed a significant main effect of tone, indicating that stimuli with tones 1 and 2 were longer than stimuli with tone 4 ($B = -0.09$, $SE = 0.03$, $t = -3.56$, $p < .001$), and a significant main effect of stimulus type, indicating that words were longer than analogs ($B = 0.09$, $SE = 0.01$, $t = 6.32$, $p < .001$). Importantly, however, the interaction between tone and stimulus type failed to reach significance ($B = 0.02$, $SE = 0.03$, $t = 0.59$, $p = .56$), indicating that these effects did not vary systematically by one another.

Table 1

*Means (Standard Deviations in Parentheses) for Durations of Mandarin Word and Analogs by Tone*

<table>
<thead>
<tr>
<th></th>
<th>Word</th>
<th>Analog</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tone 1</td>
<td>0.62 (0.11)</td>
<td>0.53 (0.09)</td>
</tr>
<tr>
<td>Tone 2</td>
<td>0.59 (0.12)</td>
<td>0.51 (0.08)</td>
</tr>
<tr>
<td>Tone 4</td>
<td>0.52 (0.09)</td>
<td>0.42 (0.08)</td>
</tr>
</tbody>
</table>
Videos used to teach Mandarin words during the learning phase of the experiment were recorded and then edited with Adobe Premiere Pro (Version 14.0). For videos used in the no motion and dot conditions, a female native Mandarin speaker stood still for approximately two seconds while being video recorded. All videos were recorded to capture just the torso of the speaker, so as to exclude the face, making the discrepancy between the genders of the video and audio speakers less obvious. In gesture condition videos, the native Mandarin speaker was video recorded as she produced gestures conveying the pitch contours of each of the three Mandarin lexical tones (from left-to-right, mimicking how a student might see a teacher produce these gestures during learning).

Dot videos were created in post-production by superimposing a red moving dot over the speaker’s gestures in gesture videos. Underlying gesture videos were then deleted, and moving dots were subsequently superimposed over videos from the no motion condition. This resulted in videos showing the native Mandarin speaker, motionless, as a dot traced the pitch contour of a tone via motion identical to that of pitch gesture.

Audio tracks of all videos were deleted in post-production and replaced with audio recordings of Mandarin words and analogs, which were aligned with motion onset in gesture and dot videos, resulting in motion coinciding with sound.

To create incongruent gesture and dot videos, congruent videos from these conditions were replaced with videos from the same condition corresponding to the two other tones. To keep audio-visual tone mappings consistent for each participant, two incongruent versions were created (see Table 2), which were counterbalanced across participants.
Table 2

Incongruent Motion Versions

<table>
<thead>
<tr>
<th>Auditory Stimulus Tone</th>
<th>Version 1 Visual Stimulus Tone</th>
<th>Version 2 Visual Stimulus Tone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tone 1</td>
<td>Tone 2</td>
<td>Tone 4</td>
</tr>
<tr>
<td>Tone 2</td>
<td>Tone 4</td>
<td>Tone 1</td>
</tr>
<tr>
<td>Tone 4</td>
<td>Tone 1</td>
<td>Tone 2</td>
</tr>
</tbody>
</table>

Procedure

All experimental procedures were conducted remotely. Prior to beginning the experiment, participants were presented with an electronic informed consent form. Participants providing consent to participate then followed a link to the online study, which was hosted on Pavlovia (https://pavlovia.org) and completed on devices with an external keyboard. Before beginning the experiment, participants were advised to wear headphones and to ensure that they were in a quiet, non-distracting environment.

In this experiment, participants completed two sequences: a Mandarin word learning and an analog learning sequence. The order of these sequences was counterbalanced among participants (see Figure 1). In the gesture and dot motion conditions, the first sequence always consisted of congruent trials and the second sequence always consisted of incongruent trials to avoid transfer of incongruent tone-stimulus mappings from incongruent to congruent trials. Thus, if any transfer between trials of different congruency occurred, it was from congruent to incongruent trials, increasing the probability of Type II error for congruency. The first sequence was completed in full (i.e., pre-test, learning phase, post-test) before the second sequence began.
Each sequence (Mandarin word or analog learning) began with instructions for the pre-test. The pre-test consisted of triplets from both lists. During each trial, an auditory stimulus was heard while images depicting the pitch contours of tones 1, 2, and 4 were shown on-screen (Figure 2). Participants were instructed to press the key on the keyboard corresponding to the labelled graphical representation of the tone of the auditory stimulus they heard as quickly and accurately as possible. Once they had done so, the experiment immediately proceeded to the next trial. Each auditory stimulus was presented once. To increase the salience of differences in tone, auditory stimuli in each triplet were always presented consecutively. The order of auditory stimuli within triplets was randomized for each participant, as was the order of triplets.
After completing the pre-test, participants continued to the learning portion of the sequence, in which Mandarin word or analog triplets from one list (counterbalanced across participants) were presented across two blocks (see Figure 1). Participants were randomly assigned to one of three learning conditions before beginning the experiment: gesture \( (n = 25) \), dot \( (n = 27) \), or no motion \( (n = 12) \). In the gesture condition, participants saw videos of a native Mandarin speaker using hand gestures to depict the pitch contours of tones (Figure 3A). In the dot condition, participants saw the same videos as in the gesture condition, but each hand gesture was replaced with a small red dot that moved across the screen to convey the pitch contours of tones (Figure 3B). In the no-motion condition, participants saw videos of the native Mandarin speaker standing motionless (Figure 3C).
Depending on the sequence (Mandarin word or analog learning), participants heard seven triplets of Mandarin words or analogs while watching videos from their assigned condition (gesture, dot, no-motion). In each sequence, Mandarin words or analogs were presented in two learning blocks, the first of which contained four triplets and the second of which contained three triplets.

Each learning block was followed by a short quiz in which triplets of Mandarin words or analogs presented in the preceding block were played while one of the images depicting tone pitch contours from the pre-test was shown on-screen. In each trial, participants indicated whether the image on the screen depicted the tone of the Mandarin word or analog they heard by pressing a key on the keyboard corresponding to “Match” or “No Match.” Following the quiz, participants received feedback as to how many items they answered correctly or incorrectly.

Following the learning portion of the sequence, participants completed the post-test, which was identical to the pre-test in make-up and structure.

After completing the post-test of the first sequence, participants completed the second sequence (pre-test, learning, post-test). If Mandarin lexical tones were learned in the first sequence, the second sequence contained analogs, and vice versa. In the gesture or dot conditions, videos were shown from the same condition that were incongruent with the tones of
Mandarin words or analogs based on the mappings of one of two versions (see Table 1). In the no motion condition, videos in the second sequence were the same as those in the first sequence.

Following the post-test of the second sequence, participants \( n = 55 \) followed a link to complete a demographics questionnaire on Qualtrics and were then directed to the Harvard Music Lab website, where they completed a brief (~2.5 min.) tone deafness test to examine their musical pitch perception (https://www.themusiclaboratory.org/quizzes/td). This tone deafness test, which consists of pure tone sequences and yields a score ranging from 0-32, has been validated against the Montreal Battery of Amusia, a more extensive (~45 min., 72-item) test of tone deafness (Peretz et al., 2008), in 271 typically developing adults, and performance on the two tests is significantly correlated \( r = .27, p < .0001; \) Mehr et al., 2017).

**Results**

See Table 3 for descriptive statistics for tone identification accuracy by congruency, tone type, condition, and test. Because analyses of tone identification latency (reaction time) are typically based on correct trials, which varied substantially by condition (see Table 3), we do not report or discuss them in the main text, but they are available as Supplementary Material. Moreover, because models of tone identification accuracy including tone deafness test score as an additional fixed factor separate from the main independent variables of interest failed to converge due to the reduced number of participants who completed the tone deafness test in addition to the experimental task, and because the results of these non-converging models, taken at face value, indicated that tone deafness score failed to significantly predict tone identification accuracy, the models reported below do not include it, and we do not discuss it further. Likewise, because models of tone identification accuracy including trial number, item position in triplets
(1, 2, 3), and item list (learned vs. unlearned) as additional fixed factors separate from the main independent variables of interest failed to significantly predict tone identification accuracy, the models reported below do not include it, and we do not discuss it further.

**Table 3**

*Means (Standard Deviations in Parentheses) for Tone Identification Accuracy by Congruency, Tone Type, Condition, and Test*

<table>
<thead>
<tr>
<th>Congruent</th>
<th>Lexical</th>
<th></th>
<th></th>
<th></th>
<th>Analog</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dot</td>
<td>Gesture</td>
<td>None</td>
<td>Analog</td>
<td>Dot</td>
<td>Gesture</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>Pre-test</td>
<td>0.36 (0.48)</td>
<td>0.36 (0.48)</td>
<td>0.33 (0.47)</td>
<td>0.78 (0.41)</td>
<td>0.81 (0.39)</td>
<td>0.97 (0.18)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Post-test</td>
<td>0.46 (0.50)</td>
<td>0.41 (0.49)</td>
<td>0.39 (0.49)</td>
<td>0.85 (0.36)</td>
<td>0.89 (0.31)</td>
<td>0.99 (0.09)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Incongruent</td>
<td>0.36 (0.48)</td>
<td>0.32 (0.47)</td>
<td>0.33 (0.47)</td>
<td>0.90 (0.29)</td>
<td>0.90 (0.29)</td>
<td>0.97 (0.18)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Post-test</td>
<td>0.34 (0.47)</td>
<td>0.34 (0.47)</td>
<td>0.39 (0.49)</td>
<td>0.74 (0.44)</td>
<td>0.69 (0.46)</td>
<td>0.99 (0.09)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Identification accuracy for Mandarin lexical tones and their analogs was analyzed using logit mixed effects models, which model the log odds of a correct response (binary) on each trial. Because congruence varied in the pitch gesture and dot conditions but not the no motion condition, main analyses of accuracy and latency were conducted via three models, described below. In these models, all fixed effects were coded using weighted mean centered (Helmert) contrast coding. The maximal random effect structure permitting convergence was used to minimize Type I error. All models were fit in R using the `glmer()` function of the `lme4` package (Bates et al., 2015). Null hypothesis significance testing was conducted using the `lmerTest` package (Kuznetsova et al., 2017). One-sample *t*-tests were also conducted to compare cell means for accuracy data to chance (0.33); comparisons exceeded chance except where noted. To test for differences between more than two levels, Tukey HSD post-hoc tests were conducted.
using the *emmeans* package (Lenth, 2019). All data and analysis scripts are publicly available via the following link: [https://osf.io/2wd87/](https://osf.io/2wd87/).

To examine whether embodied dynamic visual stimuli (pitch gesture) and non-embodied dynamic visual stimuli (dot motion) based on the vertical conceptual metaphor of pitch elicit greater pre-test to post-test changes in tone identification accuracy than the absence of such stimuli (no motion), models including fixed effects of condition (dot, gesture, no motion), test (pre-test, post-test), and tone type (lexical, analog) with crossed random intercepts were used to analyze data from the congruent and incongruent conditions separately (in addition to data from the no motion condition, which was neutral with respect to congruency). For tones learned with congruent or neutral stimuli, we observed a significant two-way interaction of condition and tone type for the dot and gesture vs. no motion conditions (see Table 4). Simple effect analyses revealed that, although tone identification accuracy was higher for analogs than for lexical tones across conditions, there was a greater difference between lexical tones and analogs in the no motion condition ($B = 4.46, SE = 0.34, t = 13.27, p < .001$) than in the dot condition ($B = 2.49, SE = 0.46, t = 5.43, p < .001$) and the gesture condition ($B = 2.65, SE = 0.52, t = 5.07, p < .001$; see Figure 4). Moreover, identification accuracy for lexical tones learned with no motion failed to exceed chance ($t = 1.63; p = .10$), whereas identification accuracy exceeded chance for lexical tones learned with congruent dot motion ($t = 5.39; p < 0.001$) and congruent pitch gesture ($t = 3.64; p < 0.001$). We also observed a significant two-way interaction of test and tone type (see Table 3). Simple effect analyses revealed that, although tone identification accuracy increased from pre-test to post-test for both lexical tones and analogs, this increase was greater for analogs ($B = 0.71, SE = 0.13, t = 5.44, p < .001$) than for lexical tones ($B = 0.35, SE = 0.08, t = 4.22, p <
Moreover, identification accuracy for lexical tones failed to exceed chance at pre-test ($t = 1.80; p = .07$), whereas it exceeded chance at post-test ($t = 7.35; p < .001$).

**Table 4**

*Fixed Effect (Top) and Variance Estimates (Bottom) in Log Odds for Multi-Level Logit Model of Tone Identification Accuracy for Congruent and Neutral Stimuli (Observations = 5,376)*

<table>
<thead>
<tr>
<th>Fixed effect</th>
<th>Coefficient</th>
<th>SE</th>
<th>Wald</th>
<th>$z$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.07</td>
<td>0.15</td>
<td>7.25</td>
<td>&lt; .001***</td>
<td></td>
</tr>
<tr>
<td>Test</td>
<td>0.58</td>
<td>0.10</td>
<td>5.92</td>
<td>&lt; .001***</td>
<td></td>
</tr>
<tr>
<td>Condition (dot vs. gesture and no motion)</td>
<td>0.10</td>
<td>0.31</td>
<td>0.32</td>
<td>.75</td>
<td></td>
</tr>
<tr>
<td>Condition (dot and gesture vs. no motion)</td>
<td>1.00</td>
<td>0.40</td>
<td>2.49</td>
<td>.01*</td>
<td></td>
</tr>
<tr>
<td>Tone type</td>
<td>3.02</td>
<td>0.29</td>
<td>10.38</td>
<td>&lt; .001***</td>
<td></td>
</tr>
<tr>
<td>Test x condition (dot vs. gesture and no motion)</td>
<td>0.03</td>
<td>0.16</td>
<td>0.16</td>
<td>.87</td>
<td></td>
</tr>
<tr>
<td>Test x condition (dot and gesture vs. no motion)</td>
<td>0.33</td>
<td>0.40</td>
<td>0.83</td>
<td>.41</td>
<td></td>
</tr>
<tr>
<td>Test x tone type</td>
<td>0.47</td>
<td>0.20</td>
<td>2.40</td>
<td>.02*</td>
<td></td>
</tr>
<tr>
<td>Condition (dot vs. gesture and no motion) x tone type</td>
<td>0.20</td>
<td>0.63</td>
<td>0.32</td>
<td>.75</td>
<td></td>
</tr>
<tr>
<td>Condition (dot and gesture vs. no motion) x tone type</td>
<td>2.35</td>
<td>0.80</td>
<td>2.92</td>
<td>.003**</td>
<td></td>
</tr>
<tr>
<td>Test x condition (dot vs. gesture and no motion) x tone type</td>
<td>0.48</td>
<td>0.33</td>
<td>1.49</td>
<td>.14</td>
<td></td>
</tr>
<tr>
<td>Test x condition (dot and gesture vs. no motion) x tone type</td>
<td>0.81</td>
<td>0.80</td>
<td>1.02</td>
<td>.31</td>
<td></td>
</tr>
</tbody>
</table>

**Random effect**

<table>
<thead>
<tr>
<th></th>
<th>$s^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant</td>
<td>1.07</td>
</tr>
<tr>
<td>Item</td>
<td>0.11</td>
</tr>
</tbody>
</table>

**Figure 4**

*Tone Identification Accuracy for Congruent Dot and Gesture Conditions and No Motion*

*Condition in Pre-Test and Post-Test by Tone Type (Dashed Lines Represent Chance)*
For tones learned with incongruent or neutral stimuli, we observed a significant three-way interaction of test, condition, and tone type for dot and gesture vs. no motion (see Table 5). Simple effect analyses revealed that, for lexical tones, the interaction between test and condition failed to reach significance for both dot vs. gesture and no motion ($B = 0.14, SE = 0.17, t = 0.78, p = .43$) as well as dot and gesture vs. no motion ($B = 0.29, SE = 0.21, t = 1.43, p = .15$), as did the main effects of test ($B = 0.04, SE = 0.08, t = 0.47, p = .64$) and dot vs. gesture and no motion ($B = -0.07, SE = 0.09, t = -0.72, p = .47$) as well as dot and gesture vs. no motion ($B = 0.13, SE = 0.11, t = 1.17, p = .24$), as is evident from the similar distributions for lexical tone across conditions and tests in Figure 5. Moreover, identification accuracy for lexical tones failed to exceed chance in the dot condition at pre-test ($t = 1.23; p = .22$) as well as post-test ($t = 0.29; p = .78$), in the gesture condition at pre-test ($t = -0.29; p = .77$) as well as post-test ($t = 0.35; p = 0.72$), and in the no motion condition at pre-test ($t = 0.11; p = .91$) but not at post-test ($t = 2.16; p = .03$). For analogs, by contrast, simple effect analyses revealed that, although the interaction between test and condition failed to reach significance for dot vs. gesture and no motion ($B = -0.22, SE = 0.28, t = -0.80, p = .43$), it reached significance for dot and gesture vs. no motion ($B =$
3.07, \( SE = 0.76, t = 4.04, p < .001 \). Further examination of this effect revealed that accuracy decreased significantly from pre-test to post-test for analogs learned with incongruent dot motion (\( B = -1.54, SE = 0.20, t = -7.84, p < .001 \)) as well as incongruent pitch gesture (\( B = -1.76, SE = 0.20, t = -8.69, p < .001 \)), whereas it showed a non-significant trend towards increasing from pre-test to post-test for analogs learned with no motion (\( B = 1.41, SE = 0.80, t = 1.77, p = .08 \)), as is evident in Figure 5.

Table 5

*Fixed Effect (Top) and Variance Estimates (Bottom) in Log Odds for Multi-Level Logit Model of Tone Identification Accuracy for Incongruent and Neutral Stimuli (Observations = 5,543)*

<table>
<thead>
<tr>
<th>Fixed effect</th>
<th>Coefficient</th>
<th>SE</th>
<th>Wald</th>
<th>( p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.75</td>
<td>0.12</td>
<td>6.49</td>
<td>&lt; .001***</td>
</tr>
<tr>
<td>Test</td>
<td>-0.48</td>
<td>0.09</td>
<td>-5.03</td>
<td>&lt; .001***</td>
</tr>
<tr>
<td>Condition (dot vs. gesture and no motion)</td>
<td>-0.13</td>
<td>0.23</td>
<td>-0.54</td>
<td>.59</td>
</tr>
<tr>
<td>Condition (dot and gesture vs. no motion)</td>
<td>1.20</td>
<td>0.32</td>
<td>3.72</td>
<td>.001***</td>
</tr>
<tr>
<td>Tone type</td>
<td>3.04</td>
<td>0.22</td>
<td>13.53</td>
<td>&lt; .001***</td>
</tr>
<tr>
<td>Test x condition (dot vs. gesture and no motion)</td>
<td>-0.04</td>
<td>0.16</td>
<td>-0.23</td>
<td>.87</td>
</tr>
<tr>
<td>Test x condition (dot and gesture vs. no motion)</td>
<td>1.58</td>
<td>0.38</td>
<td>4.11</td>
<td>&lt; .001***</td>
</tr>
<tr>
<td>Test x tone type</td>
<td>-1.08</td>
<td>0.20</td>
<td>-5.48</td>
<td>&lt; .001***</td>
</tr>
<tr>
<td>Condition (dot vs. gesture and no motion) x tone type</td>
<td>-0.11</td>
<td>0.47</td>
<td>-0.24</td>
<td>.81</td>
</tr>
<tr>
<td>Condition (dot and gesture vs. no motion) x tone type</td>
<td>2.28</td>
<td>0.66</td>
<td>3.45</td>
<td>&lt; .001***</td>
</tr>
<tr>
<td>Test x condition (dot vs. gesture and no motion) x tone type</td>
<td>-0.34</td>
<td>0.33</td>
<td>-1.04</td>
<td>.30</td>
</tr>
<tr>
<td>Test x condition (dot and gesture vs. no motion) x tone type</td>
<td>2.70</td>
<td>0.81</td>
<td>3.34</td>
<td>&lt; .001***</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Random effect</th>
<th>( s^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant</td>
<td>0.78</td>
</tr>
<tr>
<td>Item</td>
<td>0.12</td>
</tr>
</tbody>
</table>

Figure 5

*Tone Identification Accuracy for Incongruent Dot and Gesture Conditions and No Motion*

*Condition in Pre-Test and Post-Test by Tone Type (Dashed Lines Represent Chance)*
Taken together, these results provide evidence that tone learning differs when visual stimuli congruent or incongruent with the vertical conceptual metaphor of pitch are present, as in the gesture and dot conditions, relative to when they are absent, as in the no motion condition. Moreover, they suggest that lexical tone and analog learning did not differ significantly between the pitch gesture and dot motion conditions. Because null hypothesis significance testing only permits rejection of the null hypothesis, however, additional evidence is needed to definitively conclude that pre-test to post-test changes do not differ significantly between the pitch gesture and dot motion conditions. A Bayes Factor approach, which quantifies support for one model over another regardless of the correctness of the models, is capable of offering such additional evidence (Morey & Rouder, 2011). Thus, we implemented this approach via the BayesFactor package (Morey & Rouder, 2018) to compare a model with fixed factors of test, tone type, congruency, and their interactions to a model with fixed factors of condition, test, tone type, congruency, and their interactions. Both models included participant as a random factor and were based on data from participants assigned only to the dot and gesture conditions. This comparison revealed that the model without condition and its interactions with other factors was 93,636.41 ±
4.5% times more likely than the model with condition and its interactions with other factors, providing confirmatory evidence that lexical tone and analog learning did not differ significantly between the pitch gesture and dot motion conditions.

Based on these findings, we conducted a follow-up analysis collapsing across the dot and gesture conditions on data from participants assigned only to these conditions to examine how congruency with the vertical conceptual metaphor of pitch affects tone identification accuracy, as well as to characterize the effects of visual stimuli based on this metaphor on lexical tone and non-speech analog learning. The model used for this analysis included fixed effects of congruency (congruent, incongruent), test (pre-test, post-test), and tone type (lexical, analog) with crossed random intercepts.

This analysis revealed a significant three-way interaction of congruency, test, and tone type (see Table 6). Simple effect analyses revealed that, for tones learned with congruent stimuli, the interaction between test and tone type showed a marginal trend towards significance ($B = 0.30$, $SE = 0.16$, $t = 1.85$, $p = .06$). However, this analysis revealed a main effect of test, indicating that tone identification accuracy increased from pre-test to post-test ($B = 0.52$, $SE = 0.08$, $t = 6.46$, $p < .001$), as well as a main effect of tone type, indicating that analogs were identified more accurately than in lexical tones ($B = 2.62$, $SE = 0.35$, $t = 7.47$, $p < .001$; see Figure 6, left panel). Moreover, pre-test accuracy for lexical tones learned with congruent stimuli failed to exceed chance ($t = 1.93$; $p = .05$), whereas post-test accuracy for these tones exceeded chance ($t = 7.11$; $p < .001$), indicating that learning lexical tones via congruent dot motion and pitch gesture increased increased tone identification accuracy from below to above chance. By contrast, the interaction between test and tone type reached significance for tones learned with incongruent stimuli ($B = -1.59$, $SE = 0.17$, $t = -9.57$, $p < .001$). Further examination of this effect
revealed that, although accuracy did not change significantly from pre-test to post-test for lexical tones ($B = -0.02, SE = 0.09, t = -0.22, p = .83$), it decreased significantly for analogs ($B = -1.66, SE = 0.14, t = -11.70, p < .001$; see Figure 6, right panel). Moreover, accuracy for lexical tones learned with incongruent stimuli failed to exceed chance at pre-test ($t = 0.68; p = .50$) as well as post-test ($t = 0.45; p = .65$), indicating that learning lexical tones via incongruent dot motion and pitch gesture failed to increase tone identification accuracy to above chance. Together, these findings are consistent with our hypothesis that visual stimuli congruent with the vertical conceptual metaphor of pitch would enhance tone learning, whereas pitch gesture and dot motion incongruent with this conceptual metaphor would hinder tone learning. Furthermore, they are consistent with our hypothesis that visual stimuli based on the vertical conceptual metaphor of pitch would not elicit significantly different effects or would elicit amplified effects on learning and would be the same in direction for non-speech analogs relative to lexical tones.

### Table 6

**Fixed Effect (Top) and Variance Estimates (Bottom) in Log Odds for Multi-Level Logit Model of Tone Identification Accuracy in Dot and Gesture Conditions (Observations = 8,903)**

<table>
<thead>
<tr>
<th>Fixed effect</th>
<th>Coefficient</th>
<th>SE</th>
<th>Wald z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.54</td>
<td>0.10</td>
<td>5.12</td>
<td>&lt;.001***</td>
</tr>
<tr>
<td>Test</td>
<td>-0.13</td>
<td>0.05</td>
<td>-2.48</td>
<td>.01*</td>
</tr>
<tr>
<td>Condition</td>
<td>0.02</td>
<td>0.20</td>
<td>0.11</td>
<td>.91</td>
</tr>
<tr>
<td>Congruency</td>
<td>-0.16</td>
<td>0.06</td>
<td>-2.90</td>
<td>.004**</td>
</tr>
<tr>
<td>Tone type</td>
<td>2.32</td>
<td>0.06</td>
<td>41.06</td>
<td>&lt;.001***</td>
</tr>
<tr>
<td>Test x condition</td>
<td>-0.03</td>
<td>0.11</td>
<td>-0.29</td>
<td>.77</td>
</tr>
<tr>
<td>Test x congruency</td>
<td>-1.16</td>
<td>0.11</td>
<td>-10.64</td>
<td>&lt;.001***</td>
</tr>
<tr>
<td>Condition x congruency</td>
<td>-0.18</td>
<td>0.11</td>
<td>-1.57</td>
<td>.12</td>
</tr>
<tr>
<td>Test x tone type</td>
<td>-0.60</td>
<td>0.11</td>
<td>-5.49</td>
<td>&lt;.001***</td>
</tr>
<tr>
<td>Condition x tone type</td>
<td>0.19</td>
<td>0.11</td>
<td>1.73</td>
<td>.08</td>
</tr>
<tr>
<td>Congruency x tone type</td>
<td>0.31</td>
<td>0.39</td>
<td>0.79</td>
<td>.43</td>
</tr>
<tr>
<td>Test x condition x congruency</td>
<td>-0.01</td>
<td>0.22</td>
<td>-0.06</td>
<td>.96</td>
</tr>
<tr>
<td>Test x condition x tone type</td>
<td>0.06</td>
<td>0.22</td>
<td>0.27</td>
<td>.79</td>
</tr>
</tbody>
</table>
In this experiment, we investigated the effects of conceptual metaphor and embodiment on lexical and non-speech tone learning via several interrelated research questions. First, we
examined how the presence of dynamic visual stimuli based on the vertical conceptual metaphor of pitch (pitch gesture, dot motion) affects tone learning relative to its absence (no motion). We predicted that the presence of these stimuli would elicit greater pre-test to post-test changes in tone identification accuracy than their absence based on evidence that the vertical conceptual metaphor of pitch and embodiment of it via gesture affects lexical tone acquisition (Baills et al., 2019; Bluhme & Burr, 1971; Godfroid et al., 2017; Liu et al., 2011; Morett & Chang, 2015; Zhen et al., 2019; Zheng et al., 2018) as well as pitch perception in musical stimuli (Ben-Artzi & Marks, 1995; Casasanto et al., 2003; Connell et al., 2013; Evans & Treisman, 2010; Getz & Kubovy, 2018; Godøy et al., 2006; Küssner et al., 2014; Melara & O’Brien, 1987; Mossbridge et al., 2011; Patching & Quinlan, 2002). Our results were generally inconsistent with this prediction. For tones learned with dot motion and pitch gesture congruent with the vertical conceptual metaphor of pitch, changes in identification accuracy did not differ significantly from that of tones learned with no motion. For tones learned with dot motion and pitch gesture incongruent with the vertical conceptual metaphor of pitch, by contrast, changes in identification accuracy differed significantly from that of tones learned with no motion. This was due to a difference in directionality rather than magnitude, however; pre-test to post-test accuracy decreased in the dot motion and pitch gesture conditions whereas it increased in the no motion condition. This difference for congruent vs. incongruent visual stimuli relative to the no motion condition raises the possibility that including incongruent as well as congruent dot motion and pitch gesture may have reduced the reliability of visual cues in the incongruent condition, leading to a decrease in identification accuracy from pre-test to post-test. By including only congruent items in the learning block of the first sequence and only incongruent items in the learning block of the second sequence and by keeping incongruent audio-visual tone mappings
consistent across all incongruent items for each participant, however, we attempted to minimize any such confounding effects. Thus, we believe that cross-modal tone mapping interference induced by the incongruent condition rather than inclusion of incongruent as well as congruent visual stimuli explains the difference between changes in identification accuracy for congruent vs. incongruent visual stimuli relative to no motion.

Second, we compared the effects of observing embodied (pitch gesture) and non-embodied (dot motion) dynamic visual stimuli based on the vertical conceptual metaphor of pitch on tone learning. We predicted that pre-test to post-test changes in identification accuracy would not differ significantly for tones learned while observing pitch gesture and dot motion based on evidence that observing non-embodied as well as embodied visual stimuli based on the vertical conceptual metaphor of pitch affects lexical tone learning (Baills et al., 2019; Bluhme & Burr, 1971; Godfroid et al., 2017; Liu et al., 2011; Morett & Chang, 2015; Zhen et al., 2019) as well as pitch perception in musical stimuli (Ben-Artzi & Marks, 1995; Casasanto et al., 2003; Connell et al., 2013; Evans & Treisman, 2010; Getz & Kubovy, 2018; Godøy et al., 2006; Küssner et al., 2014; Melara & O’Brien, 1987; Mossbridge et al., 2011; Patching & Quinlan, 2002). Our results were consistent with this prediction, providing convergent evidence from null hypothesis significance testing and a Bayes Factor approach that pre-test to post-test changes in identification accuracy did not differ significantly for tones learned while observing dot motion and pitch gesture. This finding suggests that the vertical conceptual metaphor of pitch, regardless of whether it is conveyed via meaningful embodied action or non-embodied motion, affects tone learning in speech as well as non-speech auditory stimuli. Notably, the experimental design of our study permits direct comparison of the effects of embodied and non-embodied visual stimuli based on the vertical conceptual metaphor of pitch on lexical tone as well as non-speech analog
LEXICAL AND NON-SPEECH TONE LEARNING

learning. Thus, our results build upon the results of previous work examining the effects of these stimuli separately on lexical as well as non-speech tone learning, providing compelling evidence that conceptual metaphor, rather than embodiment, is capable of facilitating tone learning.

Third, we examined how congruency of dynamic visual stimuli (pitch gestures, dot motion) with the vertical conceptual metaphor of pitch affects tone learning via cross-modal mapping. We predicted that learning tones while observing pitch gesture and dot motion congruent with the vertical conceptual metaphor of pitch would increase the accuracy of their identification from pre-test to post-test based on evidence that visual stimuli in which the upwards direction is associated with high pitch and the downwards direction is associated with low pitch facilitate lexical tone acquisition (Baills et al., 2019; Bluhme & Burr, 1971; Godfroid et al., 2017; Liu et al., 2011; Morett & Chang, 2015; Zhen et al., 2019). Our results were consistent with this prediction, indicating that identification accuracy increased from pre-test to post-test for tones learned with congruent dot motion and pitch gesture. This finding is consistent with previous research demonstrating beneficial effects of embodied and non-embodied stimuli congruent with the vertical conceptual metaphor of pitch on perception of pitch in lexical tone (Baills et al., 2019; Bluhme & Burr, 1971; Godfroid et al., 2017; Liu et al., 2011; Morett & Chang, 2015; Zhen et al., 2019; Zheng et al., 2018) as well as musical tone (Ben-Artzi & Marks, 1995; Casasanto et al., 2003; Connell et al., 2013; Evans & Treisman, 2010; Getz & Kubovy, 2018; Godøy et al., 2006; Küssner et al., 2014; Melara & O’Brien, 1987; Mossbridge et al., 2011; Patching & Quinlan, 2002). Conversely, we predicted that learning tones while observing pitch gesture and dot motion incongruent with the vertical conceptual metaphor of pitch would decrease the accuracy of their identification from pre-test to post-test by disrupting cross-modal mapping of visual space onto auditory pitch. Our results were partially consistent with this
prediction. With respect to tones learned with incongruent dot motion and pitch gesture, identification accuracy did not change significantly from pre-test to post-test for lexical tones, whereas it decreased significantly for non-speech analogs. Although it is unclear why dot motion and pitch gesture incongruent with tone affected lexical and non-speech tone learning differently, the lack of improvement in identification accuracy from pre-test to post-test for both types of tones provides evidence that visual stimuli incongruent with the vertical conceptual metaphor of pitch interfere with tone learning by disrupting cross-modal mapping of visual space onto auditory pitch. Given that this study is the first to investigate the effects of visual stimuli incongruent with the vertical conceptual metaphor of pitch on tone learning, it will be important to include this condition in future research to determine the robustness of these findings.

Fourth, we investigated whether conceptual metaphor and embodiment affect lexical tone and non-speech analog learning similarly. We predicted that the change in tone identification accuracy from pre-test to post-test would not differ significantly or would be amplified and would be the same in direction for non-speech analogs relative to lexical tones learned with dynamic visual stimuli based on the vertical conceptual metaphor of pitch (dot motion, pitch gesture). This prediction was based on evidence that the vertical conceptual metaphor of pitch and embodiment of it via gesture affects pitch perception in musical stimuli (Ben-Artzi & Marks, 1995; Casasanto et al., 2003; Connell et al., 2013; Evans & Treisman, 2010; Getz & Kubovy, 2018; Godøy et al., 2006; Küssner et al., 2014; Melara & O’Brien, 1987; Mossbridge et al., 2011; Patching & Quinlan, 2002) similarly to lexical tone acquisition (Baills et al., 2019; Bluhme & Burr, 1971; Godfroid et al., 2017; Liu et al., 2011; Morett & Chang, 2015; Zhen et al., 2019; Zheng et al., 2018). Our results were consistent with this prediction. For tones learned with congruent dot motion and pitch gesture, identification accuracy increased significantly from pre-
test to post-test regardless of tone type, demonstrating that the effects of these visual stimuli on learning did not differ significantly for non-speech analogs relative to lexical tones. For tones learned with incongruent dot motion and pitch gesture, identification accuracy, identification accuracy did not change significantly from pre-test to post-test for lexical tones, whereas it decreased significantly for non-speech analogs, demonstrating that the effects of these visual stimuli on learning were amplified for non-speech analogs relative to lexical tones. Moreover, for tones learned with both congruent and incongruent dot motion as well as pitch gesture, the change in tone identification accuracy from pre-test to post-test was the same in direction for non-speech analogs relative to lexical tones. Taken together, these findings are consistent with domain-general theories of speech perception (Diehl et al., 2004; Samuel, 2011), indicating that conceptual metaphor and embodiment affect tone processing in speech and music similarly. Notably, the experimental design of our study permits direct comparison of the effects of visual stimuli based on the vertical conceptual metaphor of pitch on lexical tone and non-speech analog learning. Thus, our results build upon the results of previous work examining the effects of these visual stimuli on lexical and non-speech tone learning separately, providing compelling evidence of the similarity of their effects on speech and music processing and learning.

In light of our findings, future research should address the impact of additional aspects of visual stimuli on lexical and non-speech tone learning and perception. For example, rotated (congruent) pitch gestures in the horizontal plane facilitate lexical tone learning similarly to conventional (congruent) pitch gestures in the vertical plane (Zhen et al., 2019), but it is unclear whether this is the case for comparable non-embodied motion as well as musical tone learning. Moreover, research on audiovisual correspondences shows correlations between pitch in non-speech sounds and visual dimensions other than height, including size (Evans & Treisman, 2010;
LEXICAL AND NON-SPEECH TONE LEARNING

Gallace & Spence, 2006; Mondloch & Maurer, 2004; Spector & Maurer, 2009), brightness (Marks, 1974, 1987; Martino & Marks, 1999; Mondloch & Maurer, 2004), and sharpness (Marks, 1987; Maurer et al., 2012; O’Boyle & Tarte, 1980; Parise & Spence, 2009), raising the question of whether these dimensions extend to pitch in speech. These other visual dimensions may influence lexical and non-lexical tone learning in addition to the vertical conceptual metaphor of pitch, though some work suggests that this metaphor supercedes them in establishing audiovisual cross-modal mappings (e.g., Getz & Kubovy, 2018).

Given that differences between musicians and non-musicians (Besson et al., 2007; Chandrasekaran et al., 2009; Kraus & Chandrasekaran, 2010; Magne et al., 2006; Marques et al., 2007; Wong et al., 2007) and tonal and atonal language speakers (Bent et al., 2006; Chandrasekaran et al., 2007, 2009; Morett, 2020) have provided insight into the overlap in processing between speech and music, it will be important for future research to examine such differences in addition to pure tone pitch perception to determine whether conceptual metaphor and embodiment affect these groups similarly. Future research should also investigate how far the impact of conceptual metaphor and embodiment generalizes beyond Mandarin lexical tones and their non-speech analogs to more complex lexical tone systems (e.g., Thai, Cantonese) and their non-speech analogs. Moreover, future research should further probe the effect of embodiment on tone learning and perception by disentangling it from conceptual metaphor (e.g., via other types of gestures such as beat gestures, which convey emphasis) and differentiating between its direct and indirect effects by comparing gesture observation and production.

Although the current study provides insight into how conceptual metaphor and embodiment influence lexical and non-speech tone learning, it is not without limitations. One limitation is manipulation of congruency within participants in the dot motion and pitch gesture
conditions, which permitted transfer of congruent tone-stimulus mappings to test trials in the incongruent condition, increasing the probability of Type II error for congruency. Provided the participant sample size can be at least doubled, it would be ideal to manipulate audiovisual pitch mapping congruency between participants to avoid this potential confound in future work. Another limitation is exclusion of tone 3, which is an important part of the Mandarin tone system and is often confused with tone 2 by L2 learners (Hao, 2012, 2018). In future work, it will be important to include tone 3 to determine whether conceptual metaphor and embodiment can enhance differentiation between tones 2 and 3, evidence of which would further strengthen the case for their use as pedagogical aids to facilitate Mandarin lexical tone learning. Yet another limitation is the number of observations per condition, which fell short of the ideal minimum of 1,600 suggested by Brysbaert and Stevens (2014) to detect small to very small effects. Future work using between-participants designs should ideally increase the number of trials to meet or exceed this minimum to ensure that small to very small effects can be detected.

In conclusion, the current study provides the first evidence that embodied and non-embodied visual depictions of pitch contours congruent with the vertical conceptual metaphor of pitch enhance lexical and non-speech tone learning comparably. These findings suggest that non-embodied visual depictions, such as simple animations or perhaps even static images, are as effective as gestures in conveying the pitch contours of lexical and musical tones and can thus be used as pedagogical aids to facilitate learning of these tones in lieu of gestures. Thus, the current study illuminates the roles of both conceptual metaphor and embodiment in lexical and non-speech pitch processing, providing insight into how they can be leveraged to enhance learning of tonal languages as well as musical tone.
References


https://doi.org/10.1017/S0272263118000074

https://doi.org/10.18637/jss.v067.i01

https://doi.org/10.3758/BF03208371

https://doi.org/10.1037/0096-1523.32.1.97


https://doi.org/10.2307/455326


https://doi.org/10.1016/j.brainres.2006.10.064


https://doi.org/10.3389/fpsyg.2017.00297


Morey, R. D., & Rouder, J. N. (2018). *BayesFactor: Computation of Bayes Factors for common designs* (0.9.12-4.2) [Computer software]. https://CRAN.R-project.org/package=BayesFactor


Appendix A

Mandarin Word Triplets Used in Experiment

List 1

<table>
<thead>
<tr>
<th>Tone 1</th>
<th>Tone 2</th>
<th>Tone 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>cai1</td>
<td>cai2</td>
<td>cai4</td>
</tr>
<tr>
<td>ba1</td>
<td>ba2</td>
<td>ba4</td>
</tr>
<tr>
<td>ying1</td>
<td>ying2</td>
<td>ying4</td>
</tr>
<tr>
<td>mo1</td>
<td>mo2</td>
<td>mo4</td>
</tr>
<tr>
<td>piao1</td>
<td>piao2</td>
<td>piao4</td>
</tr>
<tr>
<td>chi1</td>
<td>chi2</td>
<td>chi4</td>
</tr>
<tr>
<td>fu1</td>
<td>fu2</td>
<td>fu4</td>
</tr>
</tbody>
</table>

List 2

<table>
<thead>
<tr>
<th>Tone 1</th>
<th>Tone 2</th>
<th>Tone 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>zhai1</td>
<td>zhai2</td>
<td>zhai4</td>
</tr>
<tr>
<td>hua1</td>
<td>hua2</td>
<td>hua4</td>
</tr>
<tr>
<td>xing1</td>
<td>xing2</td>
<td>xing4</td>
</tr>
<tr>
<td>guo1</td>
<td>guo2</td>
<td>guo4</td>
</tr>
<tr>
<td>nao1</td>
<td>nao2</td>
<td>nao4</td>
</tr>
<tr>
<td>ci1</td>
<td>ci2</td>
<td>ci4</td>
</tr>
<tr>
<td>tu1</td>
<td>tu2</td>
<td>tu4</td>
</tr>
</tbody>
</table>