COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 46 (2022) e13083

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13083

Embodied Space-pitch Associations are Shaped by Language

Judith Holler,[a,b] Linda Drijvers,[a,b] Afrooz Rafiee,[c] Asifa Majid[c,d]

[a]*Donders Institute for Brain, Cognition & Behaviors, Radboud University*
[b]*Language & Cognition and Neurobiology of Language Departments, Max Planck Institute for Psycholinguistics*
[c]*Canter for Language Studies, Radboud University*
[d]*Department of Psychology, University of York*

Received 25 February 2021; received in revised form 29 November 2021; accepted 2 December 2021

Abstract

Height-pitch associations are claimed to be universal and independent of language, but this claim remains controversial. The present study sheds new light on this debate with a multimodal analysis of individual sound and melody descriptions obtained in an interactive communication paradigm with speakers of Dutch and Farsi. The findings reveal that, in contrast to Dutch speakers, Farsi speakers do not use a height-pitch metaphor consistently in speech. Both Dutch and Farsi speakers' co-speech gestures did reveal a mapping of higher pitches to higher space and lower pitches to lower space, and this gesture space-pitch mapping tended to co-occur with corresponding spatial words (high-low). However, this mapping was much weaker in Farsi speakers than Dutch speakers. This suggests that cross-linguistic differences shape the conceptualization of pitch and further calls into question the universality of height-pitch associations.

Keywords: pitch; space; height-pitch metaphor; gesture; cross-linguistic; universality

1. Introduction

Sound and space are closely connected in language and thought, but the exact nature of this relationship is disputed. Adults and children from the earliest ages respond to

Correspondence should be sent to Judith Holler, Donders Institute for Brain, Cognition & Behaviour, Radboud University, P.O.Box 9010, 6500 GL, Nijmegen, The Netherlands. E-mail: judith.holler@mpi.nl

correspondences between auditory pitch and various spatial dimensions (Dolscheid, Hunnius, Casasanto, & Majid, 2014; Marks, Hammeal, Bornstein, & Smith, 1987; Möhring, Ramsook, Hirsh-Pasek, Golinkoff, & Newcombe, 2016; Roffler & Butler, 1968; Rusconi, Kwan, Giordano, Umilta, & Butterworth, 2006; Starr & Srinivasan, 2018; Walker et al., 2010, 2018)—as do non-human animals, specifically dogs (Korzeniowska, Root-Gutteridge, Simner, & Reby, 2019)—suggesting there is an inherent connection between these domains. Consistent with this, across the world's languages, spatial metaphors for pitch are ubiquitous (Eitan & Timmers, 2010; Majid et al., 2018). Accordingly, it has been claimed that height-pitch associations are universal (Evans & Treisman, 2010; Pratt, 1930; Stumpf, 1883): Even people without a height-pitch metaphor in their language display height-pitch associations in non-linguistic tasks (Parkinson, Kohler, Sievers, & Wheatley, 2012). This mapping may reflect an evolutionary adaptation to real-world auditory scene statistics since higher pitches are more likely to appear higher in space, and the human ear appears to be shaped to pick out exactly this co-occurrence (Parise, Knorre, & Ernst, 2014).

On the other hand, there is evidence inconsistent with the proposed universality of height-pitch associations. Dolscheid, Shayan, Majid, and Casasanto (2013) compared Dutch and Farsi speakers in a pitch-space correspondence task: Participants heard a musical tone while they saw an irrelevant line on the screen displayed at different heights and at a prompt had to sing it back. While Dutch speakers' reproductions were influenced by the irrelevant height-spatial stimulus, Farsi speakers were not. That is, Farsi speakers showed no evidence of a height-pitch association. Other studies also find attenuated height-pitch associations in speakers of languages without high-low metaphors, including Spanish, Catalan, and Turkish (Dolscheid, Çelik, Erkan, Küntay, & Majid, 2020; Fernandez-Prieto, Spence, Pons, & Navarra, 2017). While Spanish and Catalan descriptors of auditory pitch do not convey spatial information at all, Farsi and Turkish both use a different spatial metaphor instead—high tones are *thin*, and low tones are *thick* (Shayan, Ozturk, & Sicoli, 2011).

In fact, Farsi speakers display a range of strategies when describing pitch. High pitches are described as *thin*, but also *sharp, tranquil, mild, delicate, weak*, and an abstract term with no spatial connotation *zeer*; low pitches are described as *thick, tall/high, strong*, and another non-spatial sound term *bam* (Shayan et al., 2011). Incredibly, high spatial language has been elicited for low tones—thus suggesting a reverse height-pitch mapping to the proposed universal (Shayan et al., 2011). This makes Farsi an interesting language to consider for further study. Do Farsi speakers activate a spatial template when they listen to and talk about pitch? If so, is there evidence for activation of the universal high-low height-pitch association, or do they show evidence of an opposite or different spatial mapping?

In order to investigate this, we used a novel research paradigm in this domain utilizing language and co-speech gestures as a window into people's thinking-for-speaking about sounds. Co-speech gestures are the spontaneous movements of the hands and arms that accompany speech and often depict semantic information (iconic and metaphoric gestures; Goldin-Meadow, 2003; McNeill, 1992). While the information encoded in co-speech gesture can be largely redundant with the information in speech, gestures also depict semantic information that complements the verbal modality (McNeill, 1992) as has been shown for concrete (Holler & Beattie, 2002, 2003; Melinger & Levelt, 2004) and more abstract matters, such as

pain (e.g., Rowbotham, Holler, Lloyd, & Wearden, 2014). That is, co-speech gestures often depict information that is semantically closely related to but different from the information in speech. This means an investigation of co-speech gestures can provide additional insights into a thought, which cannot be gleaned from studying linguistic descriptions alone.

Moreover, co-speech gestures are external representations of visuo-spatial mental imagery (Hostetter & Alibali, 2008, 2019; Kita & Özyürek, 2003; McNeill, 1992) relating to both concrete and metaphoric concepts as noted above (Cienki & Müller, 2008; McNeill, 1992; Mittelberg & Waugh, 2009; Sweetser, 1997). Due to their visuo-spatial nature, co-speech gestures are particularly well-suited to depict spatial relations (Graham & Argyle, 1975; Holler & Beattie, 2002, 2003). This extends to spatial metaphors and cross-domain metaphorical mappings as can be seen in co-speech gestures representing time as space, for example (Casasanto & Jasmin, 2012; Gu, Zheng, & Swerts, 2019; Núñez & Cooperrider, 2013; Núñez & Sweetser, 2006). So there is good reason to expect co-speech gestures to represent auditory sounds in terms of spatial features that in turn reflect their cognitive representation. In particular, high-low conceptual metaphors of the auditory pitch should elicit co-speech gestures that depict higher notes higher in space and lower notes lower in space.

Previous studies suggest there may be an influence of the location of gestures in space (vertical axis) onto the perception of pitch height (Baills, Suárez-González, González-Fuente, & Prieto, 2019; Connell, Cai, & Holler, 2013; Kelly, Bailey, & Hirata, 2017; Yuan, González-Fuente, Baills, & Prieto, 2019), and kinesthetic information from bodily movements appear to have similar effects (Hostetter, Dandar, Shimko, & Grogan, 2019), providing some justification for the idea that such associations may also underpin gesture production. Similarly, imitating scripted gestures can influence the production of the pitch in second language learning (Baills et al., 2019; for more modest effects, see Zheng, Hirata, & Kelly, 2018,). When asked to communicate about auditory stimuli using non-verbal vocal expressions and gestures, the pitch is often mapped to vertical space in gestural movements (Lemaitre et al., 2017), and a similar mapping appears between eyebrow position and pitch height when singing (Huron, Dahl, & Johnson, 2009). However, no published research to date has investigated the connection between pitch and space for spontaneous gesture production in the context of linguistic descriptions of sound,[1] nor compared the potential gestural mapping of pitch onto space cross-linguistically.

In the present study, we, therefore, investigate whether speakers' gestures reveal a high-low spatial conceptualization of auditory pitch while eliciting linguistic descriptions. In particular, we ask if Farsi speakers activate a "universal" high-low spatial template for auditory pitch by examining space-pitch mappings in (a) their linguistic descriptions, and (b) their co-speech gestures. Critically, we test whether gestures for pitch map on to vertical space even if Farsi speakers' linguistic descriptions do not. To elucidate the relation between speech and gesture further, we compare Farsi to Dutch speakers whose language features a clear height-pitch metaphor. In terms of the linguistic descriptions, Dutch speakers should produce high-low linguistic metaphors, while Farsi speakers should produce a more varied set of linguistic descriptions in line with past research (e.g., Shayan et al., 2011).

In short, this study provides an original multimodal analysis of pitch descriptions that crucially informs the debate about the universality of height-pitch associations, on the one hand,

and the interplay between language, gesture, and mental imagery, on the other. If we find cross-cultural differences in space-pitch associations in gesture, this would call into question claims about the universality of height-pitch associations. At the very least, if Farsi speakers' linguistic descriptions do not show evidence of a strong, consistent mapping between height and pitch, and there are differences in height-pitch mappings in gesture, this would suggest language plays a critical role in modulating the mental representations underpinning thinking-for-speaking about sound.

2. Methods

2.1. Participants

Thirty native Dutch speakers (23F, 7M) and 30 native Farsi speakers (16F, 14M) took part in the experiment. The average age was 27 years for Dutch speakers ($SD = 6.56$, range 19 to 52 years) and 30 years for Farsi speakers ($SD = 5.94$, range 23 to 52 years). None of the participants reported motor problems or language impairments. None of the Dutch speakers were experienced in Farsi, and none of the Farsi speakers spoke Dutch. While the Dutch participants were all bilingual in English, none of the Farsi speakers had strong English skills or extensive English language training.

The experiment was conducted at Radboud University Nijmegen, the Netherlands, and in Shiraz, Iran (in a quiet location in the home or workplace). Participants were paid for participation. Ethical approval was obtained from the Social Science Faculty Ethics Committee at Radboud University. All participants gave written informed consent and were debriefed after participation.

2.2. Materials

Stimuli consisted of 16 short audio clips, each containing between one and four tones. Each tone measured 1000 ms in length when used as an individual sound stimulus, and 600 ms when used as part of a sound sequence. The Hz for each tone was chosen randomly (and ranged between 220 and 1046 Hz) as was their combination into sequences. The stimuli were created using Adobe Audition CS6.

Two different sets of stimuli were created, one for Describer A (Set 1), one for Describer B (Set 2). Half of the participants in each role (Describer A/B) listened to the stimuli running from one tone to four tones, while the other half listened to the stimuli in the reverse order. In addition, for each set of stimuli given to participants in the Describer role, a corresponding Listener stimulus set was created. The Listener stimuli were identical in terms of the number of sounds played in each trial, but they could either match or mismatch the Describer's sounds in pitch (see Table 1).

2.3. Procedure

Participants took part in pairs and sat on chairs facing each other. Each participant was assigned a tablet with the auditory stimuli and a set of headphones. They were told they would

Table 1
Stimuli used in experiment

Stimulus Set	Number of Tones	Letter Code	Hz	Listener Stimulus	Match/ Mismatch
Set 1	1	A	220	A	M
	1	B	440	J	MM
	2	C	385–659	C	M
	2	D	1046–659	D	M
	3	E	385–659–1046	F	MM
	3	F	308–792–616	M	MM
	4	G	440–1012—528–748	G	M
	4	H	484–616–836–968	P	MM
Set 2	1	I	308	I	M
	1	J	352	B	MM
	2	K	352–704	K	M
	2	L	704–352	L	M
	3	M	1046–659–385	E	MM
	3	N	616–792–308	N	M
	4	O	968–863–616–484	O	M
	4	P	748–528–1012–440	H	MM

individually listen to sounds (or short melodies) at the same time. The Describer would then describe what they heard to the Listener, who then had to indicate whether the sound/melody they had heard was the same or different. Describers were asked to give as much detail as possible about what they heard. Listeners were instructed not to reveal their answers to the Describers, but to record them by circling either "same" or "different" for each trial on an answer sheet provided by the experimenter.

To ensure that both participants were listening to the correct corresponding files, they were asked to announce the file number of the recording at the beginning of each trial. The 16 trials were presented in two blocks of eight, preceded by two practice trials. All participants played the role of Describer and Listener. Roles were swapped after 8 trials, with a short break of about five minutes between blocks. The experiment was filmed using two cameras with integrated, sensitive microphones placed in either corner of the experiment room such that there would be a frontal/lateral perspective of each participant to allow for both auditory and visual analysis.

2.4. Coding

All video recordings were synchronized (per pair) and then imported into ELAN (versions 4.9.3-5.3) (https://archive.mpi.nl/tla/elan) for coding verbal and gestural behavior.

2.5. Verbal descriptions

We coded the Describers' linguistic data. The linguistic analyses focused on spontaneous descriptions of pitch quality; therefore, source descriptions (e.g., "*it was a truck's horn"*) were

excluded from the analysis. We also restricted our analysis to descriptions of single tones. For multi-tone stimuli, only references to individual tones—not melodies—were analyzed. For example, the following three single-tone descriptions were included in the analysis "*the first sound is high, the second one lower, and the third one is high again*," whereas descriptions that referred to multiple tones at once (e.g., "*the first three are low"*) or to the melodies in their entirety (e.g., "*it goes up"*) were excluded since they do not allow for a one-to-one mapping between description and pitch of the stimulus. This led to the exclusion of 138 out of 731 sound descriptions for Dutch and 216 out of 1124 for Farsi.

All sound descriptions were then classified according to whether they referred to a spatial metaphor (e.g., high-low, thin-thick, small-big, extent, amount, volumetric), another type of metaphor (e.g., hot-cold, hard-soft), or were non-metaphorical descriptions. The focus of the present analyses is on the spatial metaphors only. A binary judgement was made regarding all linguistic descriptions categorized as a high-low metaphor (i.e., high vs. low). Intensifiers (e.g., *very* high/*very* low) and comparatives (*higher/lower*) were not distinguished. Negatives (e.g., "*it was not high"*) were coded as "*[NEG] high"* and categorized accordingly for the analysis.

2.6. Co-speech gestures

2.6.1. Gesture identification

All hand movements accompanying sound descriptions and iconically depicting semantic aspects of the sounds or deictically referring to them in space were identified.

2.6.2. Gesture space

For all gestures identified as part of step one, we coded: (a) where the gesture was performed in absolute gesture space (coded as lower, center, or upper gesture space, based on the gesture space diagram in Fig. 1, loosely adapted from McNeill (1992), and (b) where the gesture was located relative to the preceding sound-depicting gesture (coded as lower, same, or higher). For the relative space coding, by definition, single sound stimuli descriptions could not be coded, and neither could the descriptions of the first sounds in a melody.

2.6.3. Gesture shape

All gestures identified as part of step one were also coded for shape. Two broad shape categories were created, "spatial" and "volumetric." Spatial gesture shapes were those that served to locate a sound in gesture space, such as by pointing to a location in space with the index finger or the whole hand, or by holding a flat hand in a specific spatial location, palm typically facing down. Volumetric gesture shapes focused on depicting the volumetric nature of a sound, typically through depicting holding, that is, containing a sound within the hand or between the fingertips. These could be implicated in conceptualizations of sound as big-small or thin-thick. Any gestures not fitting these categories were categorized as "other" and excluded from further analyses.

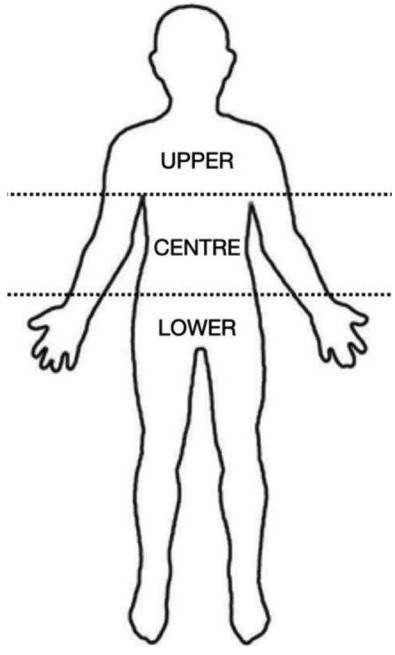

Fig. 1. Illustration of gesture space categories for the variable absolute space.[2]

2.6.4. Reliability analyses

A native speaker of Farsi (XX) coded all the Farsi speakers' gestures accompanying the descriptions of individual sounds. Two native Dutch speakers (XX and XX) coded all Dutch speakers' gestures accompanying the descriptions of individual sounds. An assistant with Dutch and basic Farsi knowledge (XX) checked the gesture coding for both languages. A second independent coder (XX) annotated the data from three Farsi speakers and three Dutch speakers for the occurrence of gestures, constituting 11.8% of the total number of gestures in the dataset ($n = 115$) to establish reliability. For gesture identification, this yielded a raw agreement of 99.12%. We observed a modified Cohen's Kappa of .89 for absolute gesture space, a modified Cohen's Kappa of .93 for relative gesture space, and a modified Cohen's Kappa of .81 for gesture shape, indicating high agreement on all dimensions (Landis & Koch, 1977).

2.7. Comprehension data

Listeners' responses were analyzed for the number of correct responses out of the total number of responses given, that is, whether the sounds they heard were the same as or different from the Describer's.

2.8. Statistical analyses

All statistical analyses were carried out using *R* (version 4.0.1, R Core Team, 2020). Following Majid et al. (2018), we used Simpson's Diversity Index (Simpson, 1949) to assess

differences in the overall naming agreement of auditory stimuli between Dutch and Farsi speakers. Differences in the use of the high-low metaphor between Dutch and Farsi speakers were assessed by fitting a generalized linear mixed model using the *lme4* package (Bates, Maechler, Bolker, & Walker, 2015). We included *Language* as a fixed effect, and random intercepts for *Participant* and *Item*, as well as random slopes for *Language* by *Item* and *Language* by *Participant*. Correlations were assessed by calculating Spearman's rho between (1) linguistic metaphor (high/low, as well as sharp-blunt for Farsi) and pitch level (ranked levels from 220 to 1048 Hz); (b) gesture in absolute space (lower, center, upper) as well as in relative space (lower, same, higher) and pitch level (ranked levels from 220 to 1048 Hz), and between (c gesture in absolute space (lower, center, upper) and linguistic metaphor (high/low), gestures in relative space (lower, same, higher) and linguistic metaphor (high/low). For each correlation, we calculated a bootstrapped confidence interval (1000 replications) to assess differences in the strength of association per group. Correlations were then statistically compared using the *cocor* package (v. 1.3, Diedenhofen & Musch, 2015). Finally, we compared differences in accuracy by fitting a generalized linear mixed model with *Language* as a fixed effect, and random intercepts for *Participant* and *Item*. Adding random slopes for *Language* by *Item* and *Language* by *Participant* led to convergence issues and were therefore omitted from the models.

The raw data and code for the statistical analysis scripts are available through the anonymized link at https://osf.io/8xtvs/?view_only = 38dd77eafbff4816a6ae9a775dccc52e.

3. Results

3.1. Verbal descriptions

We first compared the descriptions of pitch in Dutch and Farsi. Overall, Farsi speakers produced more descriptions ($n = 908$) than Dutch speakers ($n = 593$) for the same stimuli. As found in previous studies (Shayan et al., 2011), Farsi speakers produced a larger variety of responses than Dutch speakers too (see Table 2), suggesting they were entertaining multiple construals of the auditory stimuli. To quantify this, we assessed how much participants diverged when describing stimuli of the same pitch by measuring naming agreement within each group. Dutch speakers had higher naming agreement than Farsi speakers, $t(33) = 4.04$, $p < .001$, confirming the qualitative pattern.

Both groups primarily used spatial metaphors (Dutch 84.65%; Farsi 58.81%), although this pattern was considerably more pronounced for Dutch than Farsi speakers. Within the category of spatial metaphors, Dutch speakers predominantly relied on a high-low metaphor (81.28%). Contrary to the findings of Shayan et al. (2011), Farsi speakers in this study also used high-low metaphors (36.12%)—but they used this metaphor inconsistently. They also used sharp-blunt metaphors (13.66%), with thickness metaphors only accounting for 1.54% of responses.

Although Dutch and Farsi speakers both relied on a high-low metaphor, Dutch speakers used the high-low metaphor significantly more often than Farsi speakers ($\beta = 3.62$, $SE = 0.73$, $z = 4.96$, $p < .001$). Importantly, while Dutch speakers followed the universal principle,

Table 2
Types of linguistic descriptions used by Dutch and Farsi speakers with a percentage distribution of co-speech gestures; numbers in brackets refer to *n*

		Linguistic Description Types	Linguistic Description	Gestures Across Description Types	Linguistic Descriptions with Gesture
Dutch	**Spatial metaphor**	Overall	**84.65 (502)**	**86.49 (288)**	**57.37**
		High-low	81.28 (482)	83.18 (277)	57.47
		Small-big	0.17 (1)	0.00 (0)	0.00
		Extent	2.70 (16)	2.40 (8)	50.00
		Amount	0.38 (2)	0.60 (2)	100.00
		Crooked	0.17 (1)	0.33 (1)	100.00
	Other metaphor		**6.58 (39)**	**6.31 (21)**	**53.85**
	No metaphor		**8.77 (52)**	**7.21 (24)**	**46.15**
Farsi	**Spatial metaphor**	Overall	**58.81 (534)**	**62.20 (395)**	**73.97**
		High-low	36.12 (328)	39.69 (252)	76.83
		Thin-thick	1.54 (14)	1.73 (11)	78.57
		Sharp-blunt	13.66 (124)	13.54 (86)	69.35
		Small-big	0.77 (7)	0.63 (4)	57.14
		Extent	4.85 (44)	4.25 (27)	61.36
		Amount	1.76 (16)	2.20 (14)	87.50
		Front-back	0.11 (1)	0.16 (1)	100.00
	Other metaphor		**1.65 (15)**	**2.20 (14)**	**93.33**
	No metaphor		**39.54 (359)**	**35.59 (226)**	**62.95**

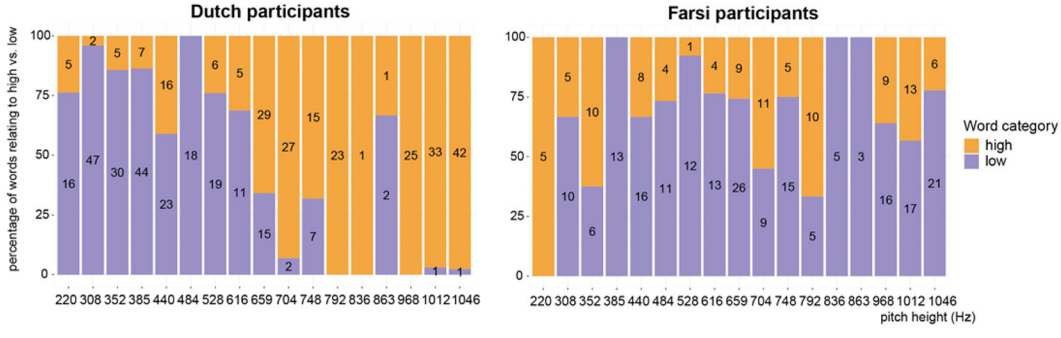

Fig. 2. Mapping of metaphoric descriptions (high-low) to pitch height for Dutch (left) and Farsi (right) speakers. Dutch speakers consistently mapped high-low metaphors to high-low pitch, but Farsi speakers did not. In all panels, numbers in the stacked bar refer to *n*.

mapping words that relate to "high" space to high pitch and words that relate to "low" space to low pitch (r_s = .68, 95% CI [0.619, 0.728], p = < .001); Farsi speakers did not (r_s = –.01, 95% CI [–0.138, 0.099], p = .80). Farsi speakers used words that relate to high and low inconsistently across pitches (see Fig. 2). That is, Farsi speakers' speech did not reveal a consistent high-low metaphor for auditory pitch. An exploratory analysis revealed a weak but statistically significant relation between "blunt" words to low pitch, and "sharp" words to

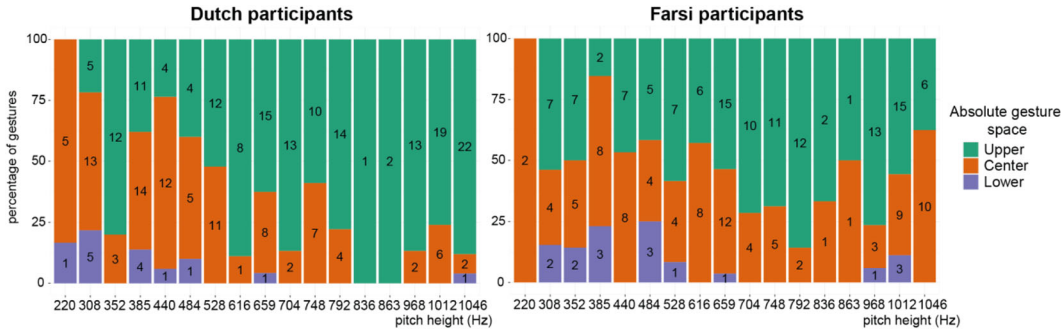

Fig. 3. Stacked bar plots with the percentage of gestures produced in the three categories of absolute gesture space related to pitch height. Dutch speakers (left panel) showed a strong mapping of absolute gesture space to pitch height. This pattern was less strong for Farsi speakers (right panel). Numbers in stacked bars refer to *n*.

higher pitch in Farsi instead ($r_s = .22$, 95% CI [0.06, 0.351], $p = .02$; see Fig. 2 but note the small *n* for the use of "blunt"). Dutch speakers never used this metaphor. No other specific spatial metaphor revealed a significant association with pitch (see Supplementary Fig. S1).

3.2. Co-speech gestures

Both Dutch and Farsi speakers often used gestures accompanying a spatial metaphor in their verbal descriptions (see Table 2). Notably, Dutch speakers consistently mapped absolute gesture space onto pitch height; gestures in lower absolute space were primarily produced for lower pitch sounds, and gestures in upper absolute space were primarily produced for higher pitch sounds ($r_s = .404$, 95% CI [0.304, 0.508], $p = < .001$). Farsi speakers also showed an association between absolute gesture space and pitch height ($r_s = .156$, 95% CI [0.026, 0.280], $p = .017$), however this was significantly weaker than for Dutch speakers ($z = 3.02$, $p = .002$, see Fig. 3).

Finally, we investigated whether Dutch and Farsi speakers consistently aligned their verbal descriptions and gestures. Both Dutch speakers and Farsi speakers primarily used gestures in upper absolute space when using words related to "high" and gestures in lower absolute space when using words related to "low" (Dutch: $r_s = .326$, 95% CI [0.210, 0.437], $p = < .001$; Farsi: $r_s = .411$, 95% CI [0.309, 0.511], $p = < .001$, with no difference between groups: $z = –1.12$, $p = .261$). Farsi speakers did not show a significant correlation between the use of gestures in upper or lower absolute space and *sharp-blunt* words ($r_s = .059$, 95% CI [–0.132, 0.337], $p = .586$; see Fig. 4).

Similar results were obtained when comparing the location of a gesture in space relative to the gesture for the preceding sound. Both Dutch and Farsi speakers primarily used gestures higher in relative space when using words related to "high" and gestures lower in relative space when using words related to "low" (Dutch: $r_s = .731$, 95% CI [0.645, 0.807], $p = < .001$; Farsi: $r_s = .521$, 95% CI [0.421, 0.618], $p = < .001$). Although this association was strong for both groups, it was significantly stronger for Dutch than Farsi speakers ($z = 3.72$, $p = < .001$). Farsi speakers did not show a significant correlation between the use of gestures

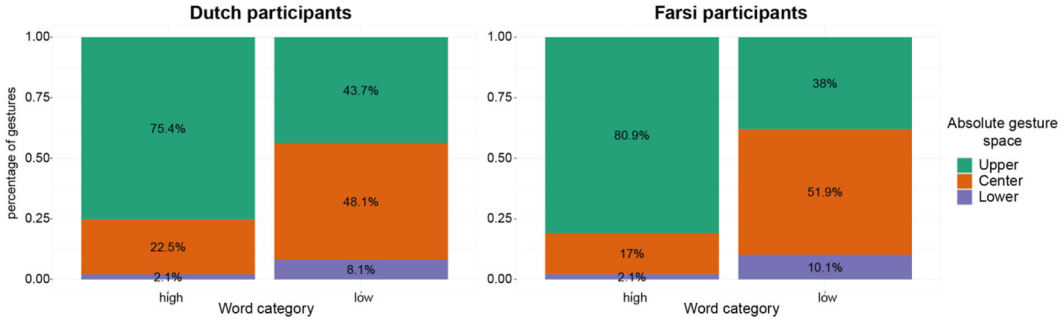

Fig. 4. Mapping of absolute gesture space (down/center/upper) to word category (high-low) for Dutch (left) and Farsi (right) speakers. Both Dutch and Farsi speakers aligned their speech with their gestures in absolute space.

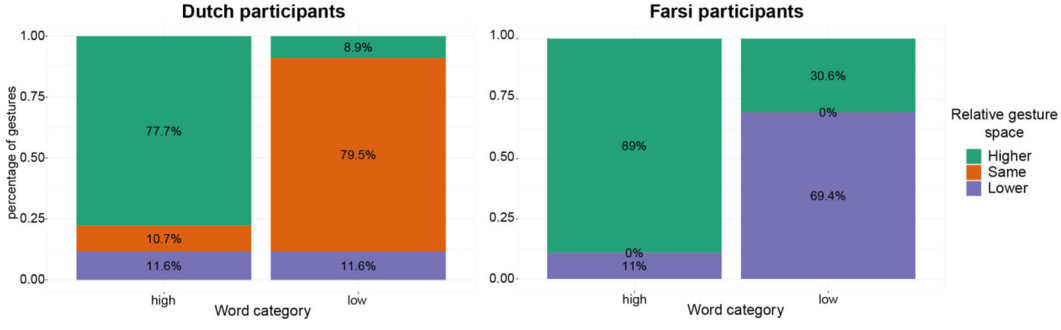

Fig. 5. Mapping of relative gesture space (lower/same/higher) to word category (high-low) for Dutch (left) and Farsi (right) speakers. Both Dutch and Farsi speakers aligned their speech with their gestures in relative space. This association was significantly stronger for Dutch than Farsi speakers.

higher or lower in relative space and sharp-blunt words ($r_s = –.001$, 95% CI [–0.196, 0.247], $p = .953$; see Fig. 5).

More generally, gesture shape did not relate to metaphor type for either Dutch or Farsi speakers (see Table S1 and Supplementary Fig. S2).

3.3. Comprehension data

Dutch listeners were significantly more accurate than Farsi listeners (Dutch 82.9% correct; Farsi 71.9% correct; $\beta = –.082$, $SE = 0.248$, $z = –3.30$, $p < .001$) at discerning whether the tones they heard matched the ones their partner had heard based on their partners' descriptions. Moreover, the presence of a gesture with spatial metaphors in speech led to higher accuracy for Dutch but not Farsi listeners ($\beta = –1.748$, $SE = 0.383$, $z = –4.56$, $p < .001$).

4. General discussion

There is ongoing debate regarding the ubiquity and universality of a correspondence between spatial height and auditory pitch across cultures (e.g., Dolscheid et al., 2013, 2020; Evans & Treisman, 2010; Korzeniowska et al., 2019; Majid et al., 2018; Parise et al., 2014; Parkinson et al., 2012; Shayan et al., 2011; Starr & Srinivasan, 2018). In this study, we used a multimodal communication paradigm in which dyads listened to individual sounds and short melodies and described what they heard to their partner. Our findings reveal novel evidence for cross-linguistic differences in the conceptualization of pitch as evidenced by both speech and co-speech gestures.

Consistent with earlier studies (e.g., Majid et al., 2018), we found that verbal descriptions of sounds were dominated by spatial metaphors; moreover, both Dutch and Farsi speakers predominantly relied on a high-low metaphor. Dutch speakers produced this metaphor more frequently than Farsi speakers, but the fact that high-low metaphors dominated Farsi speakers' responses was unexpected given previous findings (Shayan et al., 2011). In a sound description task conducted with individual speakers, Shayan et al. (2011) found the most common verbal descriptions for pitch were *nāzok* "thin" and *koloft* "thick," although on closer inspection this only accounted for 27.5% of overall responses. This was far less frequent than the thickness metaphors produced by speakers of Turkish (77%) or Zapotec (66%) in the same study. The Farsi speakers were notable for their use of a wide range of different linguistic strategies (including dedicated sound descriptors, e.g., *bam* "low pitch" that have no metaphorical connotation), which we also found here.

Shayan et al. also found that when high-low descriptors were used, they appeared in a manner inconsistent with the proposed universal height-pitch metaphorical mapping: *boland* "tall/high" was used for low pitch sounds, while *zeer* "under" was used for high pitch sounds (Shayan et al., 2011, categorized *zeer* as a non-metaphorical description, but it has a spatial meaning too). Our study found Farsi speakers applied the high-low metaphor inconsistently in speech, but we did not find evidence of a reversal of the mapping. There was essentially zero correlation between the actual pitch of a tone and the use of high-low descriptions by Farsi speakers. Farsi speakers' verbal descriptions were not entirely unsystematic, however. The next most frequent metaphor used by Farsi speakers in our study, sharp-blunt, revealed a reliable mapping between "sharp" words with high pitch and "blunt" words with a lower pitch.

In sum, while a previous study of Farsi pitch language suggested that thin-thick was the dominant metaphor, our study did not find this to be the case. Both studies, however, do show that Farsi speakers are best characterized by their wide use of diverse linguistic strategies, suggesting low codability of pitch in the language. It is unclear whether the specific differences across studies are due to dialectal differences (Tehran vs. Shiraz Farsi), rates of bilingualism (e.g., with English, a height-pitch metaphor language), differences in the paradigms employed, or something else entirely. In any case, our results clearly demonstrate that Farsi verbal descriptions do not provide strong support for the proposed universal high-low metaphor of auditory pitch.

Although Farsi speakers did not use the height-pitch metaphor consistently in speech, analysis of co-speech gesture revealed they did map higher pitches to higher space and lower pitches to lower space, and this gesture of space-pitch mapping tended to co-occur with corresponding spatial words (high-low). Nevertheless, this mapping was significantly weaker than in Dutch speakers' co-speech gestures. This suggests that non-linguistic associations between spatial height and pitch are malleable and sensitive to cross-linguistic differences in the codability of sound, contrary to what might be predicted if height-pitch mappings were the result of auditory scene statistics alone (cf., Parise et al., 2014), natural associations between pitch and affect (Huron et al., 2009), or fixed innate mappings (44h-old Italian infants appear to be sensitive to height-pitch associations; Walker et al., 2018).

Instead, our results are consistent with the proposal that the mapping between spatial height and pitch becomes more entrenched when it is bolstered by supporting metaphors in the language (Dolscheid et al., 2020). Language may not create *de novo* cross-modal mappings between sound and space, but metaphorical language appears to strengthen such associations so they are more robust. Consistent with this, Catalan and Spanish speakers without a high-low metaphor for pitch show weaker non-linguistic height-pitch associations than English speakers with a high-low metaphor (Fernandez-Prieto et al., 2017). Similarly, Turkish speakers even reverse height-pitch mappings under certain experimental conditions, which Dutch participants never do (Dolscheid et al., 2020). In the current study, we also have evidence from comprehension that further corroborates this: Farsi speakers performed significantly less accurately on the task than Dutch speakers. This suggests consistent mappings between space and pitch in verbal descriptions and gesture facilitate the communication of auditory stimuli.

On one account, the mental representations giving rise to co-speech gestures in the current study could be the result of direct perceptual experience and internalized environmental statistics, which trigger spatial imagery, entirely independent of language. However, our results suggest a more intricate interplay of language, experience, and mental imagery. We find linguistic descriptors play a critical role too. One question for the future is how different audio-spatial experiences influence one another and how they interact with linguistic encoding possibilities. Natural auditory scene statistics and affect-related associations between pitch and space can lead to conflicting cues as is the case with different musical instruments, for example. When playing the flute or piano, changes in pitch occur along the horizontal rather than vertical axis, and the pitch rises from left-to-right for the pianist but right-to-left for the flutist. This creates layers of complexities that are difficult to account for by action-based simulation accounts; moreover, such action-based influences seem fairly short-lived (Timmers & Li, 2016). How such experiences then interface with the linguistic system and its language-specific constraints remains an open question. While several gesture production models account for the interface between mental imagery and language (Hostetter & Alibali, 2019; Kita & Özyürek, 2003; McNeill, 1992), the exact details of how different factors ultimately give rise to specific multimodal expressions at the moment need further elucidation.

In sum, the present findings corroborate proposals that height-pitch associations are culturally relative by throwing new light on this issue from a multimodal communication perspective. At the same time, they help us to further refine our understanding of the interplay

between environment, action-based simulations, mental imagery, and language. Together, these findings critically inform our understanding of the cognitive processes underpinning multimodal language and thought in the communication of sound.

Notes

1 Although findings from an unpublished master's thesis (Cotroneo, 2015) do hint at a connection between pitch, space, and gesture in English speakers.
2 Original image retrieved from https://vectorified.com/body-outline-vector#body-outline-vector-23.jpg, modified in accordance with CC BY-NC 4.0 Licence.

Acknowledgements

We would like to thank Lex Pruijn and Judith Peters for help with the collection of the Dutch data, and Janna Schulze, Lex Pruijn, Judith Peters, and Ilona Plug for their contributions to coding and/or preparation of the data for processing. We would also like to thank Ludy Cilissen for editing the sound stimuli, and the Ammodo Science Foundation for financial support (Ammodo Science Award awarded to Asifa Majid).

Funding Statement: Open Access funding enabled and organized by Projekt DEAL.

WOA Institution: Max-Planck-Gesellschaft, Blended DEAL: Projekt DEAL.

References

Baills, F., Suárez-González, N., González-Fuente, S., & Prieto, P. (2019). Observing and producing pitch gestures facilitates the learning of Mandarin Chinese tones and words. *Studies in Second Language Acquisition*, *41*(1), 33–58. https://doi.org/10.1017/S0272263118000074

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Casasanto, D., & Jasmin, K. (2012). The hands of time: Temporal gestures in English speakers. *Cognitive Linguistics*, *23*(4), 643–674. https://doi.org/10.1515/cog-2012-0020

Cienki, A., & Müller, C. (Eds.). (2008). *Metaphor and gesture*. Philadelphia, PA: John Benjamins Publishing.

Connell, L., Cai, Z. G., & Holler, J. (2013). Do you see what I'm singing? Visuospatial movement biases pitch perception. *Brain and Cognition*, *81*(1), 124–130.

Cotroneo, C. (2015). From sounds to actions: how gestures depict auditory information (MPhil thesis). University of Manchester, Manchester, UK. https://www.research.manchester.ac.uk/portal/en/theses/from-sounds-to-actions-how-gestures-depict-auditoryinformation(b5a9ef23-d6a8-4ca0-8479-d80b09598a11).html.

Diedenhofen, B., & Musch, J. (2015). cocor: A comprehensive solution for the statistical comparison of correlations. *PLoS One*, *10*(4), e0121945. https://doi.org/10.1371/journal.pone.0121945

Dolscheid, S., Çelik, S., Erkan, H., Küntay, A., & Majid, A. (2020). Space-pitch associations differ in their susceptibility to language. *Cognition*, *196*, 104073. https://doi.org/10.1016/j.cognition.2019.104073

Dolscheid, S., Hunnius, S., Casasanto, D., & Majid, A. (2014). Prelinguistic infants are sensitive to space-pitch associations found across cultures. *Psychological Science*, *25*(6), 1256–1261.

Dolscheid, S., Shayan, S., Majid, A., & Casasanto, D. (2013). The thickness of musical pitch: Psychophysical evidence for linguistic relativity. *Psychological Science*, *24*(5), 613–621.

Eitan, Z., & Timmers, R. (2010). Beethoven's last piano sonata and those who follow crocodiles: Cross-domain mappings of auditory pitch in a musical context. *Cognition*, *114*(3), 405–422.

Evans, K. K., & Treisman, A. (2010). Natural cross-modal mappings between visual and auditory features. *Journal of Vision*, *10*(1), 1–12. https://doi.org/10.1167/10.1.6

Fernandez-Prieto, I., Spence, C., Pons, F., & Navarra, J. (2017). Does language influence the vertical representation of auditory pitch and loudness? *I-Perception*, *8*(3), 204166951771618. https://doi.org/10.1177/2041669517716183

Goldin-Meadow, S. (2003). *Hearing gesture: How our hands help us think*. Cambridge, MA: Belknap Press.

Graham, J. A., & Argyle, M. (1975). A cross-cultural study of the communication of extra-verbal meaning by gestures. *International Journal of Psychology*, *10*(1), 57–67.

Gu, Y., Zheng, Y., & Swerts, M. (2019). Which is in front of Chinese people, past or future? The effect of language and culture on temporal gestures and spatial conceptions of time. *Cognitive Science*, 43(12), e12804. https://doi.org/10.1111/cogs.12804

Holler, J., & Beattie, G. (2002). A micro-analytic investigation of how iconic gestures and speech represent core semantic features in talk. *Semiotica*, *1*(4), 31–69.

Holler, J., & Beattie, G. (2003). How iconic gestures and speech interact in the representation of meaning: Are both aspects really integral to the process? *Semiotica*, *146*, 81–116. https://doi.org/10.1515/semi.2003.083

Hostetter, A. B., & Alibali, M. W. (2008). Visible embodiment: Gestures as simulated action. *Psychonomic Bulletin & Review*, *15*(3), 495–514.

Hostetter, A. B., & Alibali, M. W. (2019). Gesture as simulated action: Revisiting the framework. *Psychonomic Bulletin & Review*, *26*, 721–752.

Hostetter, A. B., Dandar, C. M., Shimko, G., & Grogan, C. (2019). Reaching for the high note: Judgments of auditory pitch are affected by kinesthetic position. *Cognitive Processing*, *20*(4), 495–506. https://doi.org/10.1007/s10339-019-00929-8

Huron, D., Dahl, S., & Johnson, R. (2009). Facial expression and vocal pitch height: Evidence of an intermodal association. *Empirical Musicology Review*, *4*(3), 93–100. https://doi.org/10.18061/1811/44530

Kelly, S., Bailey, A., & Hirata, Y. (2017). Metaphoric gestures facilitate perception of intonation more than length in auditory judgments of non-native phonemic contrasts. *Collabra: Psychology*, *3*(1), 7. https://doi.org/10.1525/collabra.76

Kita, S., & Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking. *Journal of Memory and Language*, *48*(1), 16–32. https://doi.org/10.1016/S0749-596X(02)00505-3

Korzeniowska, A. T., Root-Gutteridge, H., Simner, J., & Reby, D. (2019). Audio–visual crossmodal correspondences in domestic dogs (*Canis familiaris*). *Biology Letters*, *15*(11), 20190564. https://doi.org/10.1098/rsbl.2019.0564

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.

Lemaitre, G., Scurto, H., Françoise, J., Bevilacqua, F., Houix, O., & Susini, P. (2017). Rising tones and rustling noises: Metaphors in gestural depictions of sounds. *PLoS One*, *12*(7), e0181786. https://doi.org/10.1371/journal.pone.0181786

Majid, A., Roberts, S. G., Cilissen, L., Emmorey, K., Nicodemus, B., Woll, B., LeLan, B., de Sousa, H., Cansler, B. L., Shayan, S., de Vos, C., Senft, G., Enfield, N. J., Razak, R. A., Fedden, S., Tufvesson, S., Dingemanse, M., Ozturk, O., Brown, P., Hill, C., Le Guen, O., Hirtzel, V., van Gijn, R., Sicoli, M. A., & Levinson, S. C. (2018). Differential coding of perception in the world's languages. *Proceedings of the National Academy of Sciences*, *115*(45), 11369–11376.

Marks, L. E., Hammeal, R. J., Bornstein, M. H., & Smith, L. B. (1987). Perceiving similarity and comprehending metaphor. *Monographs of the Society for Research in Child Development*, *52*(1), i–100. https://doi.org/10.2307/1166084

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago, IL: University of Chicago Press.

Melinger, A., & Levelt, W. J. M. (2004). Gesture and the communicative intention of the speaker. *Gesture*, *4*(2), 119–141.

Mittelberg, I., & Waugh, L. (2009). Metonymy first, metaphor second: A cognitive semiotic approach to multimodal figures of thought in co-speech gesture. In C. Forceville, & E. Urios-Aparisi (Eds.), *Multimodal metaphor* (pp. 329–358). New York, NY: De Gruyter Mouton.

Möhring, W., Ramsook, K. A., Hirsh-Pasek, K., Golinkoff, R. M., & Newcombe, N. S. (2016). Where music meets space: Children's sensitivity to pitch intervals is related to their mental spatial transformation skills. *Cognition*, *151*, 1–5. https://doi.org/10.1016/j.cognition.2016.02.016

Núñez, R., & Cooperrider, K. (2013). The tangle of space and time in human cognition. *Trends in Cognitive Sciences*, *17*(5), 220–229. https://doi.org/10.1016/j.tics.2013.03.008

Núñez, R., & Sweetser, E. (2006). With the future behind them: Convergent evidence from Aymara language and gesture in the crosslinguistic comparison of spatial construals of time. *Cognitive Science*, *30*(3), 401–450. https://doi.org/10.1207/s15516709cog0000_62

Parise, C. V., Knorre, K., & Ernst, M. O. (2014). Natural auditory scene statistics shapes human spatial hearing. *Proceedings of the National Academy of Sciences*, *111*(16), 6104–6108. https://doi.org/10.1073/pnas.1322705111

Parkinson, C., Kohler, P. J., Sievers, B., & Wheatley, T. (2012). Associations between auditory pitch and visual elevation do not depend on language: Evidence from a remote population. *Perception*, *41*(7), 854–861.

Pratt, C. C. (1930). The spatial character of high and low tones. *Journal of Experimental Psychology*, *13*(3), 278–285.

R Core Team. (2020). *R: A language and environment for statistical computing* (4.0.1) [Computer software]. R Foundation for Statistical Computing. http://www.R-project.org

Roffler, S. K., & Butler, R. A. (1968). Localization of tonal stimuli in the vertical plane. *The Journal of the Acoustical Society of America*, *43*(6), 1260–1266. https://doi.org/10.1121/1.1910977

Rowbotham, S., Holler, J., Lloyd, D., & Wearden, A. (2014). Handling pain: The semantic interplay of speech and co-speech hand gestures in the description of pain sensations. *Speech Communication*, *57*, 244–256.

Rusconi, E., Kwan, B., Giordano, B., Umilta, C., & Butterworth, B. (2006). Spatial representation of pitch height: The SMARC effect. *Cognition*, *99*(2), 113–129. https://doi.org/10.1016/j.cognition.2005.01.004

Shayan, S., Ozturk, O., & Sicoli, M. A. (2011). The thickness of pitch: Crossmodal metaphors in Farsi, Turkish, and Zapotec. *The Senses and Society*, *6*(1), 96–105.

Simpson, E. H. (1949). Measurement of diversity. *Nature*, *163*, 688–688.

Starr, A., & Srinivasan, M. (2018). Spatial metaphor and the development of cross-domain mappings in early childhood. *Developmental Psychology*, *54*(10), 1822–1832.

Stumpf, C. (1883). *Tonpsychologiel* (Vol. *1*). Leipzig: Hirzel.

Sweetser, E. (1997). Role and individual interpretations of change predicates. In E. Pederson, & J. Nuyts (Eds.), *Language and conceptualization* (pp. 116–136). Cambridge University Press.

Timmers, R., & Li, S. (2016). Representation of pitch in horizontal space and its dependence on musical and instrumental experience. *Psychomusicology: Music, Mind, and Brain*, *26*(2), 139–148.

Walker, P., Bremner, J. G., Lunghi, M., Dolscheid, S., D Barba, B., & Simion, F. (2018). Newborns are sensitive to the correspondence between auditory pitch and visuospatial elevation. *Developmental Psychobiology*, *60*(2), 216–223. https://doi.org/10.1002/dev.21603

Walker, P., Bremner, J. G., Mason, U., Spring, J., Mattock, K., Slater, A., & Johnson, S. P. (2010). Preverbal infants' sensitivity to synaesthetic cross-modality correspondences. *Psychological Science*, *21*(1), 21–25. https://doi.org/10.1177/0956797609354734

Yuan, C., González-Fuente, S., Baills, F., & Prieto, P. (2019). Observing pitch gestures favors the learning of Spanish intonation by Mandarin speakers. *Studies in Second Language Acquisition*, *41*(1), 5–32. https://doi.org/10.1017/S0272263117000316

Zheng, A., Hirata, Y., & Kelly, S. D. (2018). Exploring the effects of imitating hand gestures and head nods on L1 and L2 Mandarin tone production. *Journal of Speech, Language, and Hearing Research*, *61*(9), 2179–2195. https://doi.org/10.1044/2018_JSLHR-S-17-0481

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Fig. S1. Use of specific spatial metaphor (amount/extent/high-low/small-big/crooked/front-back/sharp-blunt/thin-thick) per pitch level.

Fig. S2. Use of gesture shape category (spatial or volumetric) in relation to pitch height.

Table S1 Gesture shape categories by metaphor type and by language.