Cognition 187 (2019) 178–187

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Original Articles

Manual directional gestures facilitate cross-modal perceptual learning

Anna Zhen[a,b,c,d], Stephen Van Hedger[d], Shannon Heald[d], Susan Goldin-Meadow[d], Xing Tian[a,b,c,*]

[a] Division of Arts and Sciences, New York University Shanghai, Shanghai, China
[b] Shanghai Key Laboratory of Brain Functional Genomics (Ministry of Education), School of Psychology and Cognitive Science, East China Normal University, Shanghai, China
[c] NYU-ECNU Institute of Brain and Cognitive Science at NYU Shanghai, Shanghai, China
[d] Department of Psychology, The University of Chicago, 5848 S. University Ave., Chicago, IL 60637 USA

ARTICLE INFO

Keywords:
Sensorimotor integration
Multisensory integration
Lexical tones
Gesture

ABSTRACT

Action and perception interact in complex ways to shape how we learn. In the context of language acquisition, for example, hand gestures can facilitate learning novel sound-to-meaning mappings that are critical to successfully understanding a second language. However, the mechanisms by which motor and visual information influence auditory learning are still unclear. We hypothesize that the extent to which cross-modal learning occurs is directly related to the common representational format of perceptual features across motor, visual, and auditory domains (i.e., the extent to which changes in one domain trigger similar changes in another). Furthermore, to the extent that information across modalities can be mapped onto a common representation, training in one domain may lead to learning in another domain. To test this hypothesis, we taught native English speakers Mandarin tones using directional pitch gestures. Watching or performing gestures that were congruent with pitch direction (e.g., an *up* gesture moving up, and a *down* gesture moving down, in the vertical plane) significantly enhanced tone category learning, compared to auditory-only training. Moreover, when gestures were rotated (e.g., an *up* gesture moving away from the body, and a *down* gesture moving toward the body, in the horizontal plane), performing the gestures resulted in significantly better learning, compared to watching the rotated gestures. Our results suggest that when a common representational mapping can be established between motor and sensory modalities, auditory perceptual learning is likely to be enhanced.

1. Introduction

Gestures play a vital communicative function for both speakers and listeners. For speakers, gesturing assists in the production of speech by helping individuals retrieve difficult-to-remember words from lexical memory (see Krauss, 1998; Krauss, Chen, & Chawla, 1996). For listeners, gesturing can reveal information not available in the speech signal (Driskell & Radtke, 2003; Goldin-Meadow & Alibali, 2013; Goldin-Meadow, 1999). Moreover, speakers appear to be sensitive to the benefits that gesturing bestows on listeners, as effective speakers have been shown to gesture more as message complexity increases (McNeil, Alibali, & Evans, 2000) or as background noise makes their speech harder to hear (Berger & Popelka, 1971). Beyond facilitating comprehension and production, gestures can also influence language learning. Performing gestures during learning can enhance the quantity of memorized lexical items (Zimmer, 2001), improve later recall (Masumoto et al., 2006; Schatz, Spranger, Kubik, & Knopf, 2011; Spranger, Schatz, & Knopf, 2008), lead to generalization (Wakefield, Hall, James, & Goldin-Meadow, 2018), and delay subsequent forgetting, compared to learning that is exclusively verbal (Macedonia, 2013; Tellier, 2008).

Even though previous research has shown that gestures benefit learning, the mechanisms and factors that drive this gesture-induced learning are still unclear. An influential account of how gesturing could support language learning is *multisensory learning theory* (MLT; Shams & Seitz, 2008). In this framework, gesturing benefits language learning because the addition of a concomitant motor trace during the formation of sound-to-meaning mappings leads to a more distributed and robust representation. For example, Mayer, Yildiz, Macedonia, and Kriegstein (2015) found that self-performed gestures were beneficial to foreign word learning; the correlation between this gesture benefit and neural distinctiveness (assessed through a pattern classifier) was significant in both the biological motion area of the superior temporal sulcus and the left motor cortex. These results suggest that gesturing during spoken word learning results in improved performance because the learned sound has a more distinctive, distributed representation, which could

* Corresponding author at: New York University Shanghai, 1555 Century Avenue, Room 1259, Shanghai 200122, China.
E-mail address: xing.tian@nyu.edu (X. Tian).

https://doi.org/10.1016/j.cognition.2019.03.004
Received 6 September 2018; Received in revised form 4 March 2019; Accepted 6 March 2019
Available online 14 March 2019
0010-0277/

make newly learned associations less prone to interference. However, the gestures used by Mayer et al. (2015) were related semantically to the to-be-learned word (e.g., gesturing opening a door when learning a novel word for *key*). It is therefore possible that gesture-related benefits to word learning might be restricted to cases where there is a clear relational structure between auditory and motor channels of information (e.g., Macedonia, Muller, & Friederici, 2011). In other words, not only is the *commonality of mapping* among the three modalities (auditory, visual, and motor) important for gestures to facilitate perceptual learning, but *the nature of the mapping* may also matter for learning.

Commonality of mapping—gestures that share a common representation of information with the to-be-learned stimuli – may be one of the factors that facilitate learning. For instance, Macedonia and Knosche (2011) found gesture-related benefits to word learning for iconic gestures (e.g., making an overhead, arching gesture for the word *bridge*), whereas "meaningless" gestures (e.g., touching both knees for the word *bridge*) did not result in word-learning benefits. Similarly, Kelly, Healey, Ozyurek, and Holler (2015) found participants were slower and more error-prone when gestures were incongruently (versus congruently) paired with speech. Participants had to identify gestures that illustrated an action such as pouring water into a glass, compared to speech that conveyed the same information. When speech and gestures were in direct conflict with one another, the manual and visual modalities conveyed a different representation of the information than the auditory modality. This conflict could hinder cross-modal learning simply because there is no easy mapping among the three modalities.

An important consideration in investigating the role of gesture in language learning is the level at which learning is thought to occur. In many paradigms, individuals must associate novel auditory tokens with familiar objects and concepts (e.g., associating *abiru* with *key* in Mayer et al., 2015). In order to learn the word *abiru*, the learner must be able to discriminate its perceptual features. However, for many languages, the process of learning the perceptual features that are informative for differentiating words can pose a serious challenge for non-native learners. One example is learning lexical tones, such as those found in Mandarin Chinese (see Fig. 1). In Mandarin Chinese, identical phonetic information can carry different semantic meanings depending on whether it is spoken using a high, flat pitch (Tone 1), using a rising pitch (Tone 2), using a low, dipping pitch (Tone 3), or using a falling pitch (Tone 4). For example, *ma* can either mean mother, hemp, horse, or scold/admonish in Mandarin, depending on lexical tone. For speakers of non-tonal languages to learn tonal languages such as Chinese, it is crucial for them to receive perceptual training that emphasizes the differences between lexical tones, as this kind of discrimination is a necessary first step in any ecological use of a tonal language.

Gestures have been found to be beneficial for learning the perceptual features of lexical tones (Morett & Chang, 2015), but the *nature of the mapping* involved in the process is not known. Since gestures can convey abstract ideas not available in speech (Goldin-Meadow & Alibali, 2013; Goldin-Meadow, 2003), they have the potential to provide access to relatively abstract information. On one end of the spectrum, gestures have been shown to facilitate perceptual learning when they are in complete alignment with the auditory stimuli. For example, Morett and Chang (2015) demonstrated that performing and observing iconic hand gestures that reflected the pitch changes in lexical tones supported learning Mandarin words. The pitch gestures used in their study were exact illustrations of the pitch that participants heard. It is likely that the complete alignment of pitch in manual, visual, and auditory space helped participants to learn novel Chinese words.

However, on the other end of the spectrum, gestures do not always benefit learning the perceptual features important for a given language, even when there is an alignment between the gesture and the perceptual feature. Kelly, Hirata, Manansala, and Huang (2014) assessed how well naïve listeners could learn phonemic vowel length contrasts in Japanese by asking learners to observe or produce either syllable gestures (a horizontal sweep for a long vowel and a short vertical gesture for a short vowel), or mora gestures (a short downward chopping gesture for a short vowel and two short vertical gestures for a long vowel). They found no evidence of perceptual learning despite the apparent mapping between the gesture and the to-be-learned perceptual feature. However, it is possible that the pairing between gesture and vowel length was not obvious to the learner. Intuitively for native English speakers, long and short sounds are best represented in gesture using width on a horizontal spectrum that follows the dynamics of the perceptual feature (Casasanto & Bottini, 2014). The learners in this study may not have profited from gesture because the gestures could not be easily mapped onto the to-be-learned perceptual feature.

The goal of the present study is to investigate how the *commonality and nature of mapping* impacts gesture-based improvements in perceptual learning. We hypothesized that cross-modal learning is best when it is based on a *common representational format of features across motor, visual, and auditory domains*. We used manual gestures for pitch (hand gestures that use the direction of hand and upper limb movements to visually illustrate the dynamics of pitch changes) to examine whether a common representation of an acoustic feature in the motor, visual, and auditory modalities facilitates perceptual learning. We created gestures that varied in the *ease with which they were mapped* onto the auditory signal. (1) Congruent pitch-to-gesture pairing: the trajectory and axis of the gesture could easily be mapped onto the perceptual feature (e.g., the gesture moved down in the vertical axis to represent a downward falling tone, see Fig. 2). The congruent pitch gestures pairings were aligned in features (i.e., the direction and dynamics of ascending/descending gestures mapped onto rising/falling pitch patterns) (Casasanto, Phillips, & Boroditsky, 2003). (2) Rotated pitch-to-gesture pairing: the trajectory of the gesture could be mapped onto the perceptual features of the tone, but the axis was rotated (e.g., the gesture moved toward the body in the horizontal axis to represent a downward falling tone). By rotating the pitch gestures, we removed the visual alignment between the trajectory of the gesture and the trajectory of the pitch; see Fig. 3, which displays the observer's view of the gestures and makes it clear that the gesture's trajectory is not easily mapped to the pitch in the tones (see Fig. 1). Note, however, that if the learners themselves produced the rotated gestures, they would be able to experience the gesture's trajectory and thus possibly align it to the to-be-learned

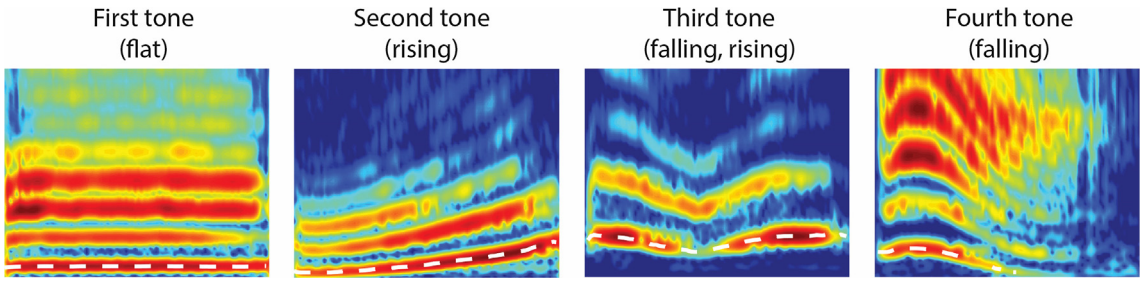

Fig. 1. Pitch contours of the four Mandarin lexical tones used in this study and displayed in spectrograms. Each tone corresponds to a pitch contour, which is displayed as flat, rising, falling-rising, and falling. The pitch contours are highlighted in the white dashed line.

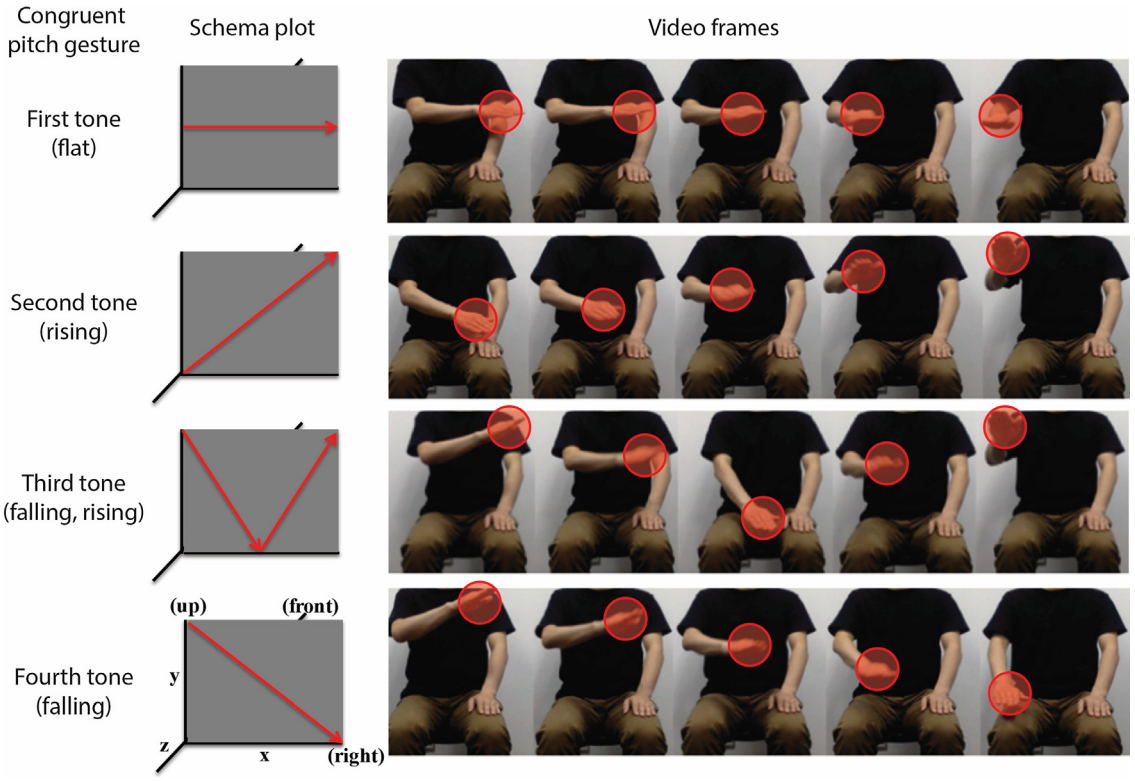

Fig. 2. Schema and video frames for the *Congruent gestures*. Congruent gestures were performed in the vertical (x-y) plane. A schema of the trajectory for each pitch is shown in a 3D plot on the left. Successive stills taken from the videos of the four congruent pitch gestures are shown on the right. Red circles highlight the trajectory of each gesture (participants did not see the circles). The trajectories mimicked the four Mandarin tone pitch contours. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

perceptual feature of the tone. The rotated pitch gestures thus allow us to address whether gestures that are not easily aligned with the auditory features of to-be-learned tone *in the visual modality* might nevertheless facilitate learning if they can be aligned with those auditory features *in the manual modality*. (3) Incongruent pitch-to-gesture pairing: neither the trajectory nor the axis of the gesture could be systematically mapped onto the perceptual feature (e.g., the gesture moved down then up in the vertical plane to represent a flat tone). These stimuli were created by randomly pairing each of the four tones displayed in Fig. 1 with one of the four gestural hand movements in Fig. 2 (excluding congruent pitch to gesture pairings).

We hypothesized that the perceptual system would align perceptual features from different modalities and more information from multiple modalities would better facilitate learning. Therefore, we approached cross-modal perceptual learning by including training that varied input from one modality (auditory), two modalities (auditory and visual), or three modalities (auditory, visual, and motor). Commonality of mapping can be influenced by input received from each of these three modalities. We compared the influence of input from modalities on perceptual learning by having training conditions that are similar, but vary on the number of modalities activated, which influences ease of mapping. We included a control where participants only heard tones during training so they would learn via input from one modality (listening to tones only), training conditions where participants received both visual and auditory cues relative to the auditory stimuli (watching gestures), and other training conditions where all three modalities received input relative to the auditory stimuli (watching and performing gestures).

We predicted that, relative to a baseline when participants listened to the lexical tones during training and saw no hand movements at all (auditory-only learning condition), congruent pitch gestures (i.e., gesture trajectory and axis are both aligned with auditory pitch) would facilitate learning. This facilitation may not depend on physically performing the gestures, as simply observing the gestures may provide sufficient information for a common representation to be created. In contrast, it is possible that horizontally rotated pitch gestures can facilitate learning when the gestures are performed, but not when they are observed (i.e., gesture trajectory is aligned with auditory pitch only when the gesture is produced by the learner, not when it is merely observed by the learner). The question is whether producing one's own gesture (and thus having kinesthetic cues to pitch trajectory) allows the learner to make enough sense of the unfamiliar visual stimulus to learn the auditory tones. Finally, we predicted that the incongruent pitch gestures would hinder learning because the mapping between pitch change and physical motion, while consistent across trials, was incongruent between modalities such that there is no semantic mapping to be made.

2. Method

2.1. Participants

108 native English speakers (32 men and 76 women) in the greater Chicago area participated in this study. Participants were between 19 and 29 years of age (M = 21.65 years, SD = 2.52). All participants reported no previous knowledge of Mandarin Chinese except for one participant, who had limited exposure to Mandarin when she was young; this participant did not perform differently from the other participants before training and was therefore included in the study. Participants were randomly assigned to one of six training conditions: *auditory only*, *perform congruent pitch gestures*, *watch congruent pitch gestures*, *perform rotated pitch gestures*, *watch rotated pitch gestures*, and *perform incongruent pitch gestures* (n = 18 for each training condition). This study was approved by IRBs at NYU Shanghai and the University of

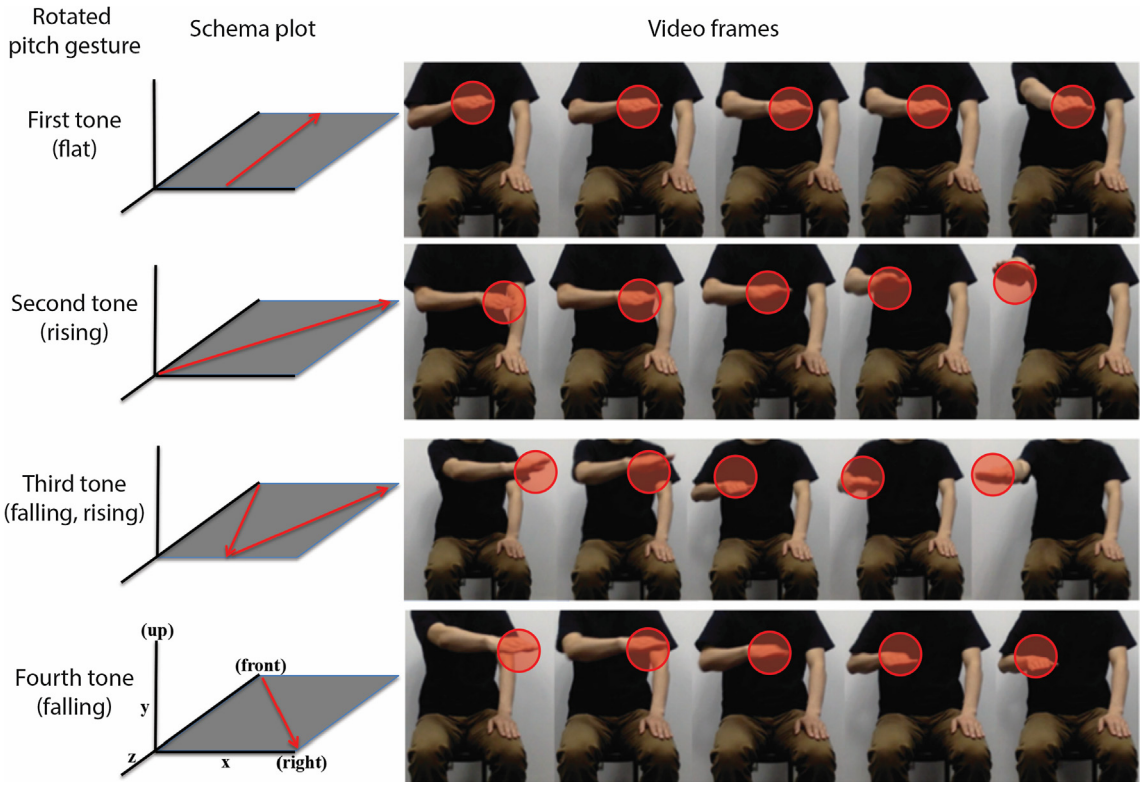

Fig. 3. Schema and video frames for the *Rotated gestures*. Rotated gestures were performed in the horizontal (x-z) plane. Each of the four congruent gestures was rotated 90 deg to form the four rotated gestures. A schema of the trajectory for each pitch gesture is shown in a 3D plot on the left. Successive stills taken from the videos of the four rotated pitch gestures are shown on the right. Red circles highlight the trajectory of each gesture (participants did not see the circles). As the dots make clear, it is difficult for an observer to map the changes in the trajectories of the gestures to the pitch dynamics in the tones. However, a participant asked to perform the four rotated gestures would experience the differences evidence in the schemas on the left.

Table 1
List of stimuli spoken by each speaker. (Each vowel and CV syllable had 4 tones.)

	Vowels	CV syllables (monosyllabic words)
Speaker One	"a": **ā, á, ǎ, à** "e": **ē, é, ě, è** "o": **ō, ó, ǒ, ò** "i": **ī, í, ǐ, ì** "u": **ū, ú, ǔ, ù** "ü": **ǖ, ǘ, ǚ, ǜ**	None
Speaker Two	"a": **ā, á, ǎ, à** "i": **ī, í, ǐ, ì** "u": **ū, ú, ǔ, ù**	< la > : **lā (**拉, pull), **lá (**剌, slash), **lǎ (**喇, woodwind instrument), **là (**辣, spicy) < li > : **lī (**哩, miles), **lí (**离, from), **lǐ (**礼, ceremony), **lì (**力, strength) < lu > : **lū (**撸,line), **lú (**庐,house), **lǔ (**卤,stew), **lù (**录, record)

Note. A total of 24 stimuli for each speaker. Speaker 1 recorded the 4 tones for each of the 6 vowels. Speaker 2 recorded the 4 tones for 3 vowels ("a", "i", and "u"), and 4 tones for CV syllables ⟨la⟩, ⟨li⟩, and ⟨lu⟩. The Chinese character and its meaning is written next to each CV syllable-tone pair.

Chicago.

The sample size of the present experiment was determined in part by previous experimental investigations of Mandarin tone learning. For example, Morett and Chang (2015), who also examined the role of pitch gestures in Mandarin tone learning, reported 19 participants per experimental group. Similarly, Wong and Perrachione (2007) investigated how well non-tonal speakers could learn Mandarin tone categories (without manual gestures) with 17 participants. Despite the present experiment deriving its sample size from these previous investigations, it should be noted that 18 participants per condition provides sufficient power only to detect large effects. For example, comparing two conditions with $n = 18$ participants in each (in a two-tailed independent samples *t*-test) would require a Cohen's d of 0.96 to reach 0.8 power. As such, we acknowledge that there may be smaller effects of interest that the present experiment is underpowered to detect.

2.2. Materials

2.2.1. Auditory stimuli

We recorded training and testing stimuli from two speakers. Speaker 1 was a male native Mandarin speaker, and Speaker 2 was a female native Mandarin speaker. Participants heard Speaker 1 in the pretest, training, and post-test. Participants heard Speaker 2 in the generalization test and in the follow-up test. The sounds from different speakers were used to test whether learning could generalize beyond the specific acoustic features of what was trained.

Six vowels in Mandarin Chinese ("a", "o", "e", "i", "u", "ü") or (/a/, /o/, /ə/, /i/, /u/, /y/) according to the International Phonetic Alphabet (IPA) were used in this study. The four tones of each of the vowels were included, which created a total of 24 stimuli for the vowels. Speaker 1 recorded these stimuli. Speaker 2 recorded the four tones of three vowels /a/, /i/, and /u/ and twelve Chinese monosyllabic words

(Table 1). The auditory words were consonant-vowel (CV) syllables that were created with the consonant ⟨l⟩ and vowels /a/, /i/, and /u/ and corresponded to actual Chinese words (Table 1). Only recordings by Speaker 1 were used in training (i.e., participants were never explicitly trained on any stimuli from Speaker 2). All auditory stimuli were 0.7 s long.

2.2.2. Visual stimuli

We recorded two sets of videos for use in the *congruent, rotated,* and *incongruent* pitch gesture conditions. The first set of videos, which were used to convey *congruent* pitch gestures (congruent gestures) (Fig. 2) were performed in the vertical plane; their trajectories were thus easily mapped onto the four tones in both trajectory and direction. Each gesture began at the gesturer's left and finished on the gesturer's right.

The second set of videos, which were used to convey *rotated* pitch gestures (rotated gestures) (Fig. 3), were recorded in the horizontal plane and had trajectories that were easily mapped onto the four tones in trajectory (left to right) but not vertical direction. The pitch direction was rotated 90 deg so an *up* gesture moves away from the body and a *down* gesture moves toward the body in the horizontal plane for *rotated* gestures. When shown these videos, participants watched trajectories that moved from their right to their left.

The *incongruent* pitch gesture (incongruent gesture) videos were created by mismatching each of the *congruent* pitch gesture videos with one of the four Mandarin Chinese tones. To make sure that there were no biases in the pairings between tones and gestures, we included all possible incongruent tone to gesture pairings, which resulted in 9 total pairings (see the 9 rows in Table 2). Each of the nine incongruent tone and gestures pairs was randomly given to two participants.

The face of Speaker 1 was not shown in the video because we wanted to remove any possible influence of mouth movements. Four tones for each vowel were dubbed into the congruent gesture videos, creating 24 videos, and into the rotated gesture videos, creating another 24 videos. There were also 24 videos for each incongruent tone to gesture pair (6 vowels × 4 tones), which created a total of 216 videos for the incongruent gestures condition. The sound tracks for all sets of videos were identical. All video clips were 3 s long: from when Speaker 1 raised his hand from his lap to start the gesture (no sound, 0.9 s), to when Speaker 1 performed the gesture and heard a 0.7 s auditory stimuli (tone) paired with the gesture (the audio started 300 ms after Speaker 1 started gesturing and ended 200 ms before he rested his hand at the height of the given pitch) and when he finished his gesture and returned his hand to his lap (no sound, 0.9 s). Moreover, the pitch dynamics in the auditory stimuli corresponded to the pitch dynamics illustrated with the gestures in the videos.

Table 2
List of the 9 incongruent matching of pitch gesture to tone for the incongruent gestures condition.

	Tone			
	Tone 1 Flat	Tone 2 Rising	Tone 3 Falling-Rising	Tone 4 Falling
Pitch Gesture	Rising	Flat	Falling	Falling-Rising
	Rising	Falling-Rising	Falling	Flat
	Rising	Falling	Flat	Falling-Rising
	Falling-Rising	Flat	Falling	Rising
	Falling-Rising	Falling	Rising	Flat
	Falling-Rising	Falling	Flat	Rising
	Falling	Flat	Rising	Falling-Rising
	Falling	Falling-Rising	Flat	Rising
	Falling	Falling-Rising	Rising	Flat

Note. Each row represents an incongruent matching of gesture to tone pair. Participants randomly received one of the 9 incongruent gestures to tone pairs. Participants saw pitch gestures that were not matched to the tone that they heard.

Table 3
Details on experimental procedures and stimuli. (Each stimulus was in four tones.)

Pretest	Training	Posttest	Generalization	Follow-Up
6 vowels	6 vowels	6 vowels	3 vowels and 3 CV syllables	3 vowels and 3 CV syllables
Speaker 1 No feedback	Speaker 1 Feedback on every trial	Speaker 1 No feedback	Speaker 2 No feedback	Speaker 2 No feedback

2.3. Procedure

Five stages (pretest, training, posttest, generalization test, and follow-up test) were included in the experiment. Pretest, training, and posttest consisted of auditory stimuli spoken by Speaker 1. There were a total of 10 blocks with 24 auditory stimuli (6 vowels × 4 tones) randomized in each block for a total of 240 trials in the pretest and 240 trials in the immediate posttest. The generalization test and the follow-up test included auditory stimuli spoken by Speaker 2. There were also a total of 10 blocks with 24 stimuli (3 vowels × 4 tones and 3 CV syllables × 4 tones) randomized in each block for a total of 240 trials in the generalization test and 240 trials in the follow-up test.

Participants first completed surveys and questionnaires on handedness (Oldfield, 1971), musical training, and knowledge of Mandarin Chinese, and also indicated their native language. Participants were then introduced to the general aspects of the Mandarin tone categorization task, and told to press buttons corresponding to Tones 1–4 (see Table 3 for the design of the experiment). These buttons were arranged horizontally on a standard computer keyboard. Stickers with the symbols "T1", "T2", "T3" and "T4" covered keys "F", "G", "H", and "J" on the keyboard. They were told to press "T1" if they thought the Mandarin tone that they heard was the first tone, "T2" if they thought it was the second tone, "T3" if they thought it was the third tone, and "T4" if they thought it was the fourth tone (see Fig. 1). Participants then completed the pretest. During the pretest, participants were simply encouraged to try their best to identify the tones they heard while looking at a fixation cross in the middle of the computer screen. No feedback was given in the pretest, and the pretest was identical across conditions. After participants completed the pretest, they were given a two-minute break.

Following the break after the completion of the pretest, participants received practice trials for the training condition to which they were randomly assigned (see Table 4). Participants saw practice videos with congruent gestures, rotated gestures, or incongruent gestures, as determined by their condition, 3 times before receiving training. There were no sounds in the practice videos. During the practice videos, participants had to perform the gestures while watching the gestures or only watch the gestures, again as determined by their condition. To mimic naturalistic observations of theses gestures as shown in classrooms or online videos, participants watched trajectories that moved from their right to their left, which is the audience perspective (see Fig. 2), and told to perform gestures from their left to their right to be consistent with the motions done by the gesturer in the video. After each gesture, participants were told to press buttons (T1, T2, T3, T4) to indicate which gesture they performed or watched. They were to press "T1" after watching and/or performing the first hand gesture, "T2" after the second hand gesture, "T3" after the third hand gesture, and "T4" after the fourth hand gesture. Participants in the auditory only condition did not receive practice since there were no gestures to become familiar with in this condition. These practice videos were the same videos that were used in training, except that there was no sound or tones in the practice videos. The practice videos were designed to help participants become familiar with the gesture-to-button association before training.

Participants received one of six types of training that had two blocks

Table 4
A description of the tasks performed in each of the 6 conditions during training.

Training conditions	Tasks in conditions			
	Listening to tone	Watching gesture	Performing gesture	Button-press
Perform congruent	✓	✓	✓	✓
Observe congruent	✓	✓		✓
Perform rotated	✓	✓ (Rotated)	✓ (Rotated)	✓
Observe rotated	✓	✓ (Rotated)		✓
Perform incongruent	✓	✓ (Incongruent)	✓ (Incongruent)	✓
Auditory only	✓			✓

Note. Participants in the *perform congruent*, *perform rotated*, and *perform incongruent* conditions saw and performed gestures. Participants in the *observe congruent* and *observe rotated* conditions only watched the gestures. Finally, participants who received the *auditory only* condition did not see or perform any gestures. Participants in all training conditions heard tones that were or were not paired with gestures and had to press T1, T2, T3, or T4 for each of the gestures they saw (for all gestures conditions) or tones that they heard (*auditory only*).

with 24 auditory or video clips randomized (6 vowels × 4 tones) in each block were included, yielding a total of 48 trials in all training conditions. In the *perform congruent pitch gestures* condition, on each trial, participants watched one of the four congruent pitch gesture videos while listening to the corresponding tone and performing the gesture they saw. They were told to press "T1" after performing the first gesture they saw in practice (flat), "T2" after performing the second gesture they saw in practice (rising), "T3" after performing the third gesture they saw in practice (falling then rising), and "T4" after performing the fourth gesture they saw in practice (falling). Gestures were explicitly associated with the buttons, whereas the tones were only implicitly associated with the buttons (via the gestures). In the *watch congruent pitch gestures* condition, participants watched one of the four congruent pitch gestures while listening to the corresponding tone. The gesture-to-button mapping was identical to the *perform congruent pitch gestures* condition. Participants in the *perform rotated pitch* gestures condition were given the same instructions as participants in the *perform congruent pitch gestures,* but saw and performed rotated pitch gesture videos rather than congruent pitch gesture videos. Participants in the *watch rotated pitch* gestures conditions were given the same instructions as participants in the *watch congruent pitch* gestures condition, but saw rotated pitch gesture videos rather than congruent pitch gesture videos. Lastly, participants in the *perform incongruent pitch gestures* condition performed and watched incongruent pitch gestures and pressed a button after they performed each gesture. Since the order in which they saw the incongruent gestures in practice matched the sequential order of the tones (see Table 2), the tone-to-button pressing was consistent with the two other performing conditions.

No video was presented in the *auditory only* condition; rather participants were given explicit feedback on the tone-to-button mapping. Given the nature of the task, one would assume that explicit feedback would enhance perceptual learning. Compared to explicit feedback from the auditory only condition, participants in the gestures condition could form an association between the gesture, tone, and button rather than just tone-to-button. They would not be memorizing the 24 different vowel and tone pairs that they hear; instead, they have only 4 gesture-to-button pairs that they need to remember, which should lighten their cognitive load, leaving room for an association between gesture and tone to form.

Participants in all conditions except the auditory only condition were presented with the video clips in training and had a maximum of 6 s to respond. Participants were given visual feedback in all learning trials on whether they had pressed the correct button ('correct' or 'incorrect' on the screen). If the participant failed to press a button in the allotted time, the words "too slow" appeared on the screen. A Samsung HMX-F90 camcorder was used to record videos of participants during training to make sure that they followed instructions and correctly performed all gestures.

Following the brief seven-minute training, participants were given a two-minute break before they had to complete the posttest, which was identical to the pretest. Immediately after the posttest, participants completed the generalization test, which followed the same procedure as the pretest and posttest, but with a different set of stimuli. Participants returned the next day to complete the follow-up test, which was identical to the generalization test.

3. Results

3.1. Accuracy

Given the design of the experiment, in which participants were randomly assigned to one of six tone-learning conditions and each completed four assessments of tone learning, we constructed a 6 (condition: auditory-only, perform-congruent, watch-congruent, perform-rotated, watch-rotated, perform-incongruent) × 4 (time: pretest, posttest, generalization test, follow-up test) mixed ANOVA with mean accuracy as the dependent variable. Training was not included in this analysis because the training stimuli differed substantially based on learning condition, unlike the other assessments in which identical tonal stimuli were presented to participants regardless of learning condition. Performance during training is reported separately in Section 3.1.5.

In this analysis[1], we observed a significant main effect of time (F (1.6, 164.6) = 149.70, $p < .001$, $\eta_p^2 = .595$), meaning performance significantly differed in at least one session from one or more of the other sessions. Post-hoc comparisons (using Bonferroni-Holm corrections) between time points showed that pretest performance was significantly worse than immediate posttest, generalization test, and follow-up test performance (all ps $< .001$, Cohen's ds = 1.00, 1.02, and 0.98, respectively). Interestingly, the immediate posttest did not appear to differ from either the generalization test or the follow-up test (ps = .73, Cohen's ds = 0.09 and 0.06, respectively), suggesting that learning generalized to a novel talker and stimuli; however, these findings should be interpreted cautiously as they rest on null findings and the present experiment is underpowered to detect small effects.

We also observed a significant main effect of learning condition (F (5, 102) = 17.20, $p < .001$, $\eta_p^2 = .457$), suggesting performance significantly differed in at least one learning condition from one or more of the other learning conditions. Post-hoc comparisons (using Bonferroni-Holm corrections) showed that the *perform-congruent, watch-congruent, and perform-rotated* conditions were all significantly more accurate than the *watch-rotated, perform-incongruent,* and *auditory-only* conditions (Table 5). The *perform-congruent, watch-congruent, and perform-rotated* conditions did not significantly differ from each other, and the *perform-incongruent* and *auditory-only* conditions did not significantly differ from each other. However, the *watch-rotated* condition was significantly more accurate than the *perform-incongruent* condition but did not differ from the *auditory-only* condition. The main effect of learning condition thus can be characterized in terms of overall superior performance for the *perform-congruent, watch-congruent,* and *perform-rotated* conditions relative to the other conditions.

The presence of a time-by-learning condition interaction (F (8.07, 164.58) = 13.08, $p < .001$, $\eta_p^2 = .391$) suggests that the differences observed among the six learning conditions were not uniform across all time points. This interaction is unpacked in the next several sections in

[1] Degrees of freedom are adjusted (Greenhouse-Geisser) to correct for sphericity.

Table 5
Pair-comparisons in possible combination of conditions at each time point of the experiment.

		Overall	Pretest	Training	Posttest	Gen. Test	Follow-Up
PC vs.							
	WC	0.97 (0.09)	2.66 (0.74)	−1.40 (0.69)	0.69 (0.25)	0.74 (0.26)	0.56 (0.21)
	PR	1.17 (0.11)	−0.52 (0.16)	−0.32 (0.12)	1.09 (0.38)	1.35 (0.43)	1.22 (0.44)
	WR	4.61 (0.44)***	0.17 (0.05)	0.80 (0.29)	4.48 (1.51)***	4.90 (1.64)***	4.21 (1.42)***
	PI	7.34 (0.71)***	1.45 (0.43)	0.32 (0.12)	7.28 (3.02)***	7.02 (2.86)***	6.88 (2.54)***
	AO	5.33 (0.51)***	0.83 (0.32)	8.06 (2.32)***	4.75 (1.78)***	5.67 (2.39)***	5.07 (2.07)***
WC vs.							
	PR	0.21 (0.02)	−3.19 (0.96)*	1.08 (0.45)	0.40 (0.12)	0.62 (0.18)	0.65 (0.21)
	WR	3.64 (0.35)**	−2.49 (0.77)	2.20 (0.93)	3.79 (1.15)**	4.16 (1.25)***	3.65 (1.10)**
	PI	6.37 (0.61)***	−1.22 (0.35)	1.73 (0.72)	6.60 (2.34)***	6.29 (2.19)***	6.32 (2.04)***
	AO	4.36 (0.42)***	−1.84 (0.66)	9.47 (2.98)***	4.06 (1.33)***	4.93 (1.76)***	4.50 (1.57)***
PR vs.							
	WR	3.44 (0.33)**	0.70 (0.24)	1.12 (0.37)	3.39 (0.98)**	3.54 (0.98)**	3.00 (0.89)*
	PI	6.17 (0.59)***	1.97 (0.64)	0.65 (0.21)	6.20 (2.08)***	5.67 (1.77)***	5.67 (1.80)***
	AO	4.16 (0.40)***	1.36 (0.60)	8.39 (2.27)***	3.66 (1.15)**	4.31 (1.37)***	3.85 (1.31)**
WR vs.							
	PI	2.73 (0.26)*	1.36 (0.60)	−0.47 (0.16)	2.80 (0.92)*	2.13 (0.70)	2.67 (0.80)
	AO	0.72 (0.07)	1.27 (0.42)	7.27 (1.98)***	0.27 (0.08)	0.77 (0.26)	0.85 (0.28)
PI vs.							
	AO	−2.01 (0.19)	−0.62 (0.25)	7.74 (2.10)***	−2.53 (0.91)	−1.36 (0.55)	−0.63 (0.36)

Note: PC = perform-congruent, WC = watch-congruent, PR = perform-rotated, WR = watch-rotated, PI = perform-incongruent, AO = auditory-only. Significance levels are adjusted using a Bonferroni-Holm correction. Numbers in parentheses represent Cohen's d effect sizes.

*** $p < .001$.

** $p < .01$.

* $p < .05$.

Table 6
Tone identification accuracy across time points for each learning condition.

	Pretest	Training	Posttest	Gen. Test	Follow-Up
PC	26.9% (9.9%)	82.6% (13.4%)***	85.1% (19.7%)***	84.5% (19.3%)***	86.5% (20.5%)***
WC	19.3% (10.6%)	90.2% (7.9%)***	79.3% (26.6%)***	78.5% (25.9%)***	81.4% (28.0%)***
PR	28.4% (8.3%)	84.4% (16.3%)***	75.6% (28.9%)***	73.4% (30.7%)***	75.6% (29.0%)***
WR	26.4% (7.8%)	78.4% (16.2%)***	46.6% (30.2%)**	44.5% (28.5%)**	48.5% (31.9%)**
PI	22.8% (9.3%)	80.9% (16.4%)***	22.3% (21.7%)	27.2% (20.8%)	24.4% (27.9%)
AO	24.5% (3.6%)	39.4% (22.8%)*	44.3% (25.9%)*	38.2% (19.3%)*	40.8% (23.7%)*

Note: PC = perform-congruent, WC = watch-congruent, PR = perform-rotated, WR = watch-rotated, PI = perform-incongruent, AO = auditory-only. Significance levels compared to a chance estimate (25%) are adjusted using a Bonferroni-Holm correction. Numbers in parentheses represent the standard deviation.

*** $p < .001$.

** $p < .01$.

* $p < .05$.

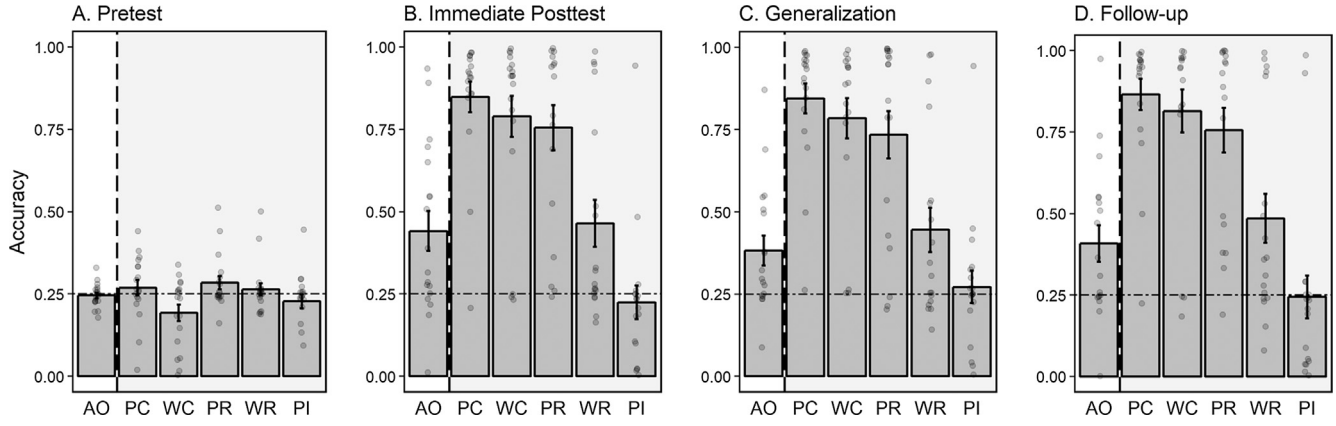

Fig. 4. Mean accuracy across the six conditions – *auditory only* (AO), *perform congruent gestures* (PC), *watch congruent gestures* (WC), *perform rotated gestures* (PR), *watch rotated gestures* (WR), and *perform incongruent gestures* (PI). The horizontal dashed line represents chance performance. Each dot represents the accuracy result of one participant. The error bars stand for +/- one standard error of the mean (SEM).

which we report post-hoc comparisons (using Bonferroni-Holm corrections) of each of the learning conditions across tonal language assessments (pretest, posttest, generalization test, and follow-up test). Performance for each of the learning conditions across each session – including training – is provided in aggregated form in Table 6 and plotted in terms of mean, standard error, and individual data points in Fig. 4.

3.1.1. Pretest

Performance was largely comparable across the learning conditions at pretest, which was expected given the random assignment of participants to learning conditions and the lack of Mandarin experience reported by the participants. Performance ranged from 19.3% in the *watch-congruent* condition to 28.4% in the *perform-rotated* condition. These two conditions were the only ones to significantly differ in the pretest after correcting for multiple comparisons (Table 5). The only condition that significantly differed from the chance estimate of 25% was the *watch-congruent* condition, which was lower than expected by chance (t (17) = −2.26, p = .037, d = 0.53); however, this significant difference does not survive corrections for multiple comparisons.

3.1.2. Posttest

Large performance differences between conditions emerged in the immediate posttest. The *perform-congruent, watch-congruent,* and *perform-rotated* gesture conditions were all significantly more accurate than the *auditory-only* condition, as well as significantly more accurate than the *watch-rotated* and *perform-incongruent* gesture conditions. The *watch-rotated* condition was significantly more accurate than the *perform-incongruent* condition. The *auditory-only* condition fell between the *watch-rotated* and *perform-incongruent* conditions and did not significantly differ from either condition; however, it was above the chance estimate of 25% (even after correcting for multiple comparisons), demonstrating significant learning. In contrast, the *perform-incongruent* gesture condition did not exceed chance performance (Table 6).

3.1.3. Generalization test

Despite experiencing novel stimuli in the generalization test, performance was remarkably similar to the posttest. Once again, the *perform-congruent, watch-congruent,* and *perform-rotated* gesture conditions were all significantly more accurate than the other conditions. The *watch-rotated* condition was nominally more accurate than the *perform-incongruent* condition, although the difference did not survive the correction for multiple comparisons. The *auditory-only* condition, which fell between the *watch-rotated* and *perform-incongruent* conditions in terms of accuracy, did not significantly differ from either condition. However, it was above the chance of 25% after correcting for multiple comparisons, which was *not* the case for the *perform-incongruent* condition.

3.1.4. Follow-up test

The follow-up test adhered to the same pattern as was observed in the posttest and the generalization test, with the *perform-congruent, watch-congruent,* and *perform-rotated* gesture conditions all performing significantly more accurately than the other conditions. The *watch-rotated* condition also significantly outperformed the *perform-incongruent* condition, with the latter condition not exceeding chance. The *auditory-only* condition fell between the *watch-rotated* and *perform-incongruent* conditions and did not statistically differ from either condition; however, it was above chance even when correcting for multiple comparisons.

3.1.5. Training

The results from the post-training assessments suggest that participants' success in learning the tones differed substantially as a function of training. However, given that training was fixed by a set number of trials (as opposed to ensuring that participants reached a certain threshold of performance), it is possible that the observed post-training differences were already present during training. Participants in the *auditory-only* condition exhibited significantly worse training performance compared to all other conditions (Table 5). This, however, is perhaps not surprising, as all other conditions involved additional practice with the gesture videos. The more critical question given the results of the post-training assessments is whether the gesture conditions significantly differed from each other in training. Training accuracy for all gesture conditions was high (ranging from 78.4% to 90.2%, Table 6), with no gesture condition significantly differing from any other gesture condition (Table 5). The fact that participants were able to establish the gesture-to-button mapping quite easily in *all* gesture conditions during training suggested that the ability to distinguish among the four gestures was not the primary cause that mediated the observed learning differences.

3.2. Response time

RTs were subject to a log transform and outlier culling (x > 3 *SD* from the mean). Similar to accuracy, we constructed a 6 (training condition: auditory-only, perform-congruent, watch-congruent, perform-rotated, watch-rotated, perform-incongruent) × 4 (time: pretest, posttest, generalization test, follow-up test) mixed ANOVA with RT as the dependent variable. In this model, we observed a significant main effect of time (F (1.9, 191.3) = 41.47, p < .001, η_p^2 = .289), suggesting RTs from at least one time point significantly differed from one or more other time points. A post-hoc test (with Bonferroni-Holm corrections) showed that *all* time points significantly differed from each other, with RTs becoming progressively faster over the course of the experiment. Unlike the analysis of accuracy, however, we did not find a significant main effect of training condition (F (5, 102) = 0.86, p = .510, η_p^2 = .041), nor did we find an interaction between time and training condition (F (9.4, 191.3) = 1.05, p = .401, η_p^2 = .049). Therefore, the observed learning differences cannot be explained by speed-accuracy tradeoff.

4. Discussion

Lexical tone learning is notoriously difficult for non-tonal speakers, as one must learn the subtle differences in pitch for semantic differentiation and thus successful communication. In the present study, we assessed how gesture can be recruited to facilitate learning patterns of pitch change representing the four lexical tone categories of Mandarin Chinese. Training that involved watching or performing gestures that were congruent with pitch direction (in the vertical plane) significantly enhanced tone category learning, relative to auditory-only training. Moreover, when gestures were rotated (onto the horizontal plane), performing but not watching gestures enhanced tone identification relative to auditory-only training. Our results are consistent with the hypothesis that a common representational mapping needs to be established between motor and sensory modalities to enhance auditory perceptual learning.

We hypothesized that the alignment between gesture and pitch would affect lexical tone learning, even in a context where the lexical tones were not the explicit focus of training. More specifically, we hypothesized that the advantage the gestures confer on lexical tone learning would be a direct consequence of the ease with which the gestures could be aligned with the tones. Gestures that clearly aligned with pitch changes in features (i.e., where the direction and dynamics of ascending/descending gestures mapped onto rising/falling pitch patterns) were hypothesized to facilitate tone learning. Other types of directional gestures, in which the relationship between gestural motion and pitch was less clear, were hypothesized to confer no benefit or even hinder auditory perceptual learning.

In the congruent pitch gesture conditions (where participants either

viewed or performed gestures that were directionally aligned with the auditory pitch change), we found robust lexical tone learning. Congruent pitch gestures were represented in vertical space, which is the dominant metaphor for describing pitch in English (e.g., Evans & Treisman, 2010), suggesting that the enhanced learning found in these conditions was driven by the transparent mapping between the motion in the gestures and the metaphoric motion in the pitches. Under multisensory learning theory, the benefit to learning in these conditions can be described in terms of shared features between auditory and visuospatial domains, which led to a more distributed and robust representation of lexical tone.

Rotated pitch gestures, in contrast, were not as transparently aligned with the changes in pitch because pitch rises and falls were represented on a horizontal plane (with rises moving away from the body and falls moving toward the body). This kind of mapping is not necessarily inconsistent with how listeners conceptualize pitch (e.g., see Eitan & Timmers, 2010), but it does represent a less commonly encountered spatial mapping for native English speakers. Moreover, the unfamiliarity with the rotated pitch gestures and their relation with pitch dynamics could be challenging for learners who observed these gestures; learners who produced the gestures themselves had the potential to gain more information. Performing rotated pitch gestures, but not observing the gestures, might thus enhance learning – which is precisely the pattern that we found. Performing rotated pitch gestures resulted in auditory perceptual learning that was comparable to performing and watching congruent pitch gesture conditions. Even though the rotated pitch gestures were not as easily mapped onto the auditory stimuli as gestures in the vertical plane, learners could form a relatively abstract mapping between the motor and auditory modalities, which facilitated perceptual learning. Simply viewing these gestures resulted in a notable reduction in performance post-training, with performance in the immediate posttest, generalization test, and next-day follow-up test not significantly differing from participants who received only auditory training.

The difference in the learning outcomes between performing rotated pitch gestures and watching rotated pitch gestures suggests that visual and auditory information cannot integrate efficiently when there is an apparent mismatch between sensory modalities. However, the mismatch had the potential to be resolved via the intermediate representation available in the motor system when action was performed, suggesting that the flexibility in motor representation linked the sensory domains. In other words, when participants performed gestures that illustrated a mapping that was not obvious visually, alignment of the tonal pitch to proprioceptive information from performance overcame the mismatch in the visual domain and enhanced cross-modal integration. The mismatch between the visual and auditory domains makes the integration less optimal, but the involvement of the motor domain increases the efficiency of cross-modal integration.

Note, however, that the performance difference between performing and watching rotated pitch gestures did *not* arise during training, where all participants were relatively successful, compared to those who did not view any gestures during training. The differences appeared only on the post-test assessments (immediate, generalization, follow-up; see Fig. 4) suggesting that merely practicing the gesture categories was not sufficient to induce learning – the gesture categories had to be meaningfully mapped onto the auditory tones. Based on the results of the rotated pitch gestures, we argue that if a learner can apprehend a representation in gesture (even if it is relatively abstract) and map it onto the auditory modality, the gesture can facilitate perceptual learning.

Results from participants who performed *incongruent* pitch gestures are also consistent with this framework. Incongruent pitch gestures were mismatched pitch-to-gesture pairings. All possible incongruent pitch and gesture combinations were included. The incongruent pitch-to-gesture pairings were consistent throughout training; thus, there was a consistent one-to-one mapping between the performed gesture and lexical tone category that could, in theory, facilitate auditory learning. However, the nature of the mapping hindered auditory learning simply because the gestural trajectories could not be transparently mapped onto the pitch changes, and thus did not meet the requirements for multisensory learning. Post-training performance was nominally worse than the *auditory only* condition and was not significantly above chance. As in the *watch rotated* pitch gesture condition, participants were able to differentiate and correctly categorize these incongruent pitch gestures during training (accurately identifying 80.9% of trials). However, given the misalignment between the gestural and auditory information, this accurate identification during training was likely at the expense of auditory learning.

To examine the robustness and transfer of the learning, participants had to extend what they had learned to novel Chinese words spoken by a speaker not encountered during training. Participants in all training conditions except the *perform incongruent pitch gestures* condition (who made no progress after training) were able to generalize their learning to new auditory stimuli. In particular, participants who performed and/or watched congruent pitch gestures or performed rotated pitch gestures demonstrated a steep increase in their ability to identify and distinguish among all of the tones in the vowels that they heard before training and generalize their learning to novel Chinese words that they heard only after training. They were able to correctly identify the tones for nearly all the vowels and words that they heard after training even though their accuracy was at chance before training.

The learning effects of training were also lasting. Participants were able to maintain their knowledge of the Mandarin tones a day after they had received training. Performance did not decline on the follow-up test for any of the groups who displayed learning after training. This finding provides support for the long-term effectiveness of brief training. Participants received 48 trials of training, which was around only seven minutes. Nonetheless, participants who had no experience with Mandarin Chinese were able to learn the four tones, and maintain and generalize that knowledge to novel Chinese words produced by a speaker that they had not heard previously. Our findings suggest that gesture is particularly useful in facilitating and maintaining learning, and leading to generalization.

This study provides consistent evidence suggesting that the *commonality* and *nature of mapping* among distinct modalities can mediate cross-modal perceptual learning. Our findings offer a novel perspective on linking motor and perceptual domains in the context of learning. However, the detailed mechanistic account needs to be further investigated. Additional experiments with different techniques are required to further illustrate how commonality and ease of mapping are implemented to facilitate the cross-modal perceptual learning. Moreover, the sample size in the present study primarily allows for large effects to be detected so subtle, but important, differences in learning outcomes between training conditions such as performing and watching congruent pitch gestures, or between watching congruent pitch gestures and performing rotated pitch gestures, might not be detected with the current sample size. Future studies should also explore factors that lead to individual differences in learning outcomes within training conditions as shown in Fig. 4. Factors such as working memory capacity, pitch discrimination abilities, and executive functions may be informative in describing the distributional nature of cross-modal learning.

In sum, our results provide new insights into the factors and mechanisms that drive cross-modal learning in the context of acquiring speech categories. This study, to our knowledge, is the first to show that there are multiple levels of mapping for perceptual features among motor and sensory modalities that drive perceptual learning. We have found that an iconic mapping between the gesture and the auditory signal is essential to facilitate learning – the arbitrary mappings in the incongruent pitch gesture condition did *not* lead to improved learning. Moreover, gesture can facilitate learning even when the mapping between movement and sound is relatively abstract – that is, when high and low sounds are associated with far and near in horizontal space, as

in the rotated pitch gestures. Importantly, rotated pitch gestures facilitate pitch learning only when participants establish the mapping between sound and movement, which they were able to do after performing but not watching the gestures. The results of this study suggest that gesturing can facilitate auditory perceptual learning as long as there is a clear mapping between the gestures and the auditory features.

Author contributions

A. Zhen and X. Tian conceived the study. A. Zhen, S. V. Hedger, S. Heald, S. Goldin-Meadow, and X. Tian designed the experiments. A. Zhen performed all experiments. A. Zhen and S. V. Hedger analyzed the data. A. Zhen, S. V. Hedger, S. Heald, S. Goldin-Meadow, and X. Tian wrote the paper. X. Tian supervised the study.

Acknowledgements

This study was supported by the National Natural Science Foundation of China 31871131, Major Program of Science and Technology Commission of Shanghai Municipality (STCSM) 17JC1404104, Program of Introducing Talents of Discipline to Universities, Base B16018, a grant from the New York University Global Seed Grants for Collaborative Research (85-65701-G0757-R4551), and the JRI Seed Grants for Research Collaboration from NYU-ECNU Institute of Brain and Cognitive Science at NYU Shanghai. We thank Jinbiao Yang for recording gesture videos and auditory stimuli, Peipei Zhang for recording auditory stimuli, Howard C. Nusbaum for commenting on the manuscript, and Haydee Marino for help with running participants in a condition.

Competing interests

The authors declare no competing interests.

Note on data availability

The data for this project and a description of it can be found at https://osf.io/qzdmj/.

http://doi.org/10.17605/OSF.IO/QZDMJ.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2019.03.004.

References

Berger, K. W., & Popelka, G. R. (1971). Extra-facial gestures in relation to speechreading. *Journal of Communication Disorders, 3*, 302–308. https://doi.org/10.1016/0021-9924(71)90036-0.

Casasanto, D., & Bottini, R. (2014). Spatial language and abstract concepts. *WIREs Cognitive Science, 5*, 139–149. https://doi.org/10.1002/wcs.1271.

Casasanto, D., Phillips, W., & Boroditsky, L. (2003). *Do we think about music in terms of space? Metaphoric representation of musical pitch. Proceedings of 25th the annual conference of the Cognitive Science Society, Boston, MA*.

Driskell, J. E., & Radtke, P. H. (2003). The effect of gesture on speech production and comprehension. *Human Factors, 45*(3), 445–454. https://doi.org/10.1518/hfes.45.3.445.27258.

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, 3*(114), 405–422. https://doi.org/10.1016/j.cognition.2009.10.013.

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

Goldin-Meadow, S. (1999). The role of gesture in communication and thinking. *Trends in Cognitive Science, 3*(11), 419–429. https://doi.org/10.1016/S1364-6613(99)01397-2.

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

Goldin-Meadow, S., & Alibali, M. W. (2013). Gesture's role in speaking, learning, and creating language. *Annual Reviews of Psychology, 64*, 257–283. https://doi.org/10.1146/annurev-psych-113011-143802.

InternationalPhoneticAlphabet.org (2016). *IPA chart with sounds. International phonetic alphabet – Promoting the study of phonetics*. International Phonetic Association Web. 24 July 2017.

Kelly, S. D., Healey, M., Ozyurek, A., & Holler, J. (2015). The processing of speech, gesture, and action during language comprehension. *Psychonomic Bulletin & Review, 22*(2), https://doi.org/10.3758/s13423-014-0681-7.

Kelly, S. D., Hirata, Y., Manansala, M., & Huang, J. (2014). Exploring the role of hand gestures in learning novel phoneme contrasts and vocabulary in a second language. *Frontiers in Psychology, 5*, 1–11. https://doi.org/10.3389/fpsyg.2014.00673.

Krauss, R. M. (1998). Why do we gesture when we speak? *Current Directions in Psychological Science, 7*(2), 54–60. https://doi.org/10.1111/1467-8721.ep13175642.

Krauss, R. M., Chen, Y., & Chawla, P. (1996). Nonverbal behavior and nonverbal communication: What do conversational hand gestures tell us. *Advances in Experimental Social Psychology, 28*, 389–450. https://doi.org/10.1016/S0065-2601(08)60241-5.

Macedonia, M. (2013). Learning a second language naturally the voice movement icon approach. *Australian Journal of Educational and Developmental Psychology, 3*(2), 102–116. https://doi.org/10.5539/jedp.v3n2p102.

Macedonia, M., & Knosche, T. R. (2011). Body in mind: How gestures empower foreign language learning. *Mind, Brain, and Education, 5*(4), 196–211. https://doi.org/10.1111/j.1751-228X.2011.01129.x.

Macedonia, M., Muller, K., & Friederici, A. D. (2011). The impact of iconic gestures on foreign language word learning and its neural substrate. *Human Brain Mapping, 32*(6), 982–998. https://doi.org/10.1002/hbm.21084.

Masumoto, K., Yamaguchi, M., Sutani, K., Tsuneto, S., Fujita, A., & Tonoike, M. (2006). Reactivation of physical motor information in the memory of action events. *Brain Research, 1101*, 102–109. https://doi.org/10.1016/j.brainres.2006.05.033.

Mayer, K. M., Yildiz, I. B., Macedonia, M., & Kriegstein, K. V. (2015). Visual and motor cortices differentially support the translation of foreign language words. *Current Biology, 25*, 530–535. https://doi.org/10.1016/j.cub.2014.11.068.

McNeil, N. M., Alibali, M. W., & Evans, J. L. (2000). The role of gesture in children's comprehension of spoken language: Now they need it, now they don't. *Journal of Nonverbal Behavior, 24*(2), 131–150. https://doi.org/10.1023/A:1006657929803.

Morett, L. M., & Chang, L. (2015). Emphasising sound and meaning: Pitch gestures enhance Mandarin lexical tone acquisition. *Language, Cognition and Neuroscience, 30*(3), 347–353. https://doi.org/10.1080/23273798.2014.923105.

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia, 9*, 97–113. https://doi.org/10.1016/0028-3932(71)90067-4.

Schatz, T. R., Spranger, T., Kubik, V., & Knopf, M. (2011). Exploring the enactment effect from an information processing view: What can we learn from serial position analyses? *Scandinavian Journal of Psychology, 52*, 509–515. https://doi.org/10.1111/j.1467-9450.2011.00893.x.

Shams, L., & Seitz, A. R. (2008). Benefits of multisensory learning. *Trends in Cognitive Sciences, 12*, 411–417. https://doi.org/10.1016/j.tics.2008.07.006.

Spranger, T., Schatz, T. R., & Knopf, M. (2008). Does action make you faster? A retrieval-based approach to the origins of the enactment effect. *Scandinavan Journal of Psychology, 49*(6), 487–495. https://doi.org/10.1111/j.1467-9450.2008.00675.x.

Tellier, M. (2008). The effect of gestures on second language memorisation by young children. *Gesture, 8*(2), 219–235. https://doi.org/10.1075/gest.8.2.06tel.

Wakefield, E. M., Hall, C., James, K. H., & Goldin-Meadow, S. (2018). Gesture for generalization: Gesture facilitates flexible learning of words for actions on objects. *Developmental Science*. https://doi.org/10.1111/desc.12656, in press.

Wong, P. C. M., & Perrachione, T. K. (2007). Learning pitch patterns in lexical identification by native English-speaking adults. *Applied Psycholinguistics, 28*(4), 565–585. https://doi.org/10.1017/S0142716407070312.

Zimmer, H. (2001). Why do Actions speak louder than words. Action memory as a variant of encodingmanipulations or the result of a specific memory system? In H. D. Zimmer, R. Cohen, J. M. J. Guynn, R. Engelkamp, & M. A. Foley (Eds.). *Memory for action: A distinct form of episodicmemory* (pp. 151–198). New York: Oxford University Press.