JSLHR

Research Article

Exploring the Effects of Imitating Hand Gestures and Head Nods on L1 and L2 Mandarin Tone Production

Annie Zheng,[a,d] Yukari Hirata,[b,d] and Spencer D. Kelly[c,d]

Purpose: This study investigated the impact of metaphoric actions—head nods and hand gestures—in producing Mandarin tones for first language (L1) and second language (L2) speakers.
Method: In 2 experiments, participants imitated videos of Mandarin tones produced under 3 conditions: (a) speech alone, (b) speech + head nods, and (c) speech + hand gestures. Fundamental frequency was recorded for both L1 (Experiment 1) and L2 (Experiment 2a) speakers, and the output of the L2 speakers was rated for tonal accuracy by 7 native Mandarin judges (Experiment 2b).
Results: Experiment 1 showed that 12 L1 speakers' fundamental frequency spectral data did not differ among the 3 conditions. In Experiment 2a, the conditions did not affect the production of 24 English speakers for the most part, but there was some evidence that hand gestures helped Tone 4. In Experiment 2b, native Mandarin judges found limited conditional differences in L2 productions, with Tone 3 showing a slight head nods benefit in a subset of "correct" L2 tokens.
Conclusion: Results suggest that metaphoric bodily actions do not influence the lowest levels of L1 speech production in a tonal language and may play a very modest role during preliminary L2 learning.

One of the daunting challenges for adults learning to speak a foreign language is to master a new set of novel speech sounds. Often, this is not just a matter of "getting the accent right." In some cases, slightly mispronouncing a word can shift between two completely different meanings. For example, in Mandarin, it can mean the difference between introducing your partner as your "husband," *lǎo gōng,* or your "laborer," *láo gōng*. Building on previous research showing that nonverbal behaviors, such as hand gestures (Gs) and head nods (HNs), are tightly integrated with the speech contours in one's native language (first language [L1]; Krahmer & Swerts, 2007; Loehr, 2007), this study explores the role that producing these bodily actions plays in the successful articulation of Mandarin Chinese, a language that poses great difficulty for native English speakers.

Characteristics of Mandarin Lexical Tones

The major acoustic correlate of Mandarin lexical tones is the fundamental frequency (F0), which is a measure of the rate at which the vocal cords open and close during phonation (Lieberman & Blumstein, 1988). As the rate of vocal cord opening and closing changes, the F0 changes, and we perceptually interpret this as a pitch change.

Mandarin makes four meaningful tone distinctions and can be characterized by F0 contour and F0 height, with each tone having its own particular F0 pattern (Jongman, Wang, Moore, & Sereno, 2006). The first and fourth tones start at about the same high F0 height, but whereas the fourth tone quickly drops and ends with the shortest duration, the first tone remains at a steady high height. The second and third tones are the longest in duration, starting at approximately the same middle F0 height, but the second tone dips very slightly at the beginning (i.e., F0 dip) before rising very quickly to a high pitch (i.e., F0 rise), whereas the third tone dips somewhere in the middle of the speech segment before rising to midheight (Jongman et al., 2006). These four tones allow for minimally contrastive syllable pairs whose meanings change simply by a change

[a]Department of Neuroscience, Washington University, St. Louis, MO
[b]Department of East Asian Languages and Literatures, Colgate University, Hamilton, NY
[c]Department of Psychological and Brain Sciences, Neuroscience Program, Colgate University, Hamilton, NY
[d]Center for Language and Brain, Colgate University, Hamilton, NY
Correspondence to Spencer D. Kelly: skelly@colgate.edu

Editor-in-Chief: Julie Liss
Editor: Bharath Chandrasekaran
Received December 26, 2017
Revision received April 1, 2018
Accepted May 7, 2018
https://doi.org/10.1044/2018_JSLHR-S-17-0481

Disclosure: *The authors have declared that no competing interests existed at the time of publication.*

in tone. For example, what distinguishes between "mother" and "to scold" is only a difference of tone, where "mother" is "ma" spoken with a high, flat pitch (first tone) and "to scold" is "ma" spoken with a high, falling tone. In this way, the F0 is a suprasegmental property of acoustics, but it is used segmentally in the sense that the tone in Mandarin rides on individual vowels, not across multiple segments such as sentential intonation.

Second Language Acquisition of Mandarin Lexical Tones

As a background for the challenges of learning proper pitch production in a second language (L2), it is useful to review literature on phoneme perception in an L1 and L2. Developing a strong phonological foundation is critical to language learning success. Phoneme identification is essential to spoken word learning in children for the L1, and that phonetic learning as early as 6 months of age is strongly correlated with later language comprehension and semantic and syntactic production skills in the second year of life (Kuhl, Conboy, Padden, Nelson, & Pruitt, 2005; Tsao, Liu, & Kuhl, 2004; Werker & Curtin, 2005; Werker, Fennell, Corcoran, & Stager, 2002). Therefore, deficits in phonetic representations can lead to poor learning outcomes at all levels of language acquisition.

Acquiring L2 phonology for adults is difficult because they are outside the "optimal period." Although infants are born with the capacity to distinguish all possible speech sound contrasts, listening experience and biological factors as we develop begin to maintain or degrade our perceptual categories as we undergo synaptic pruning (Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992; Lenneberg, 1967; Oyama, 1976; Werker & Tees, 2005). Despite this difficulty, nonnative speakers, even after the offset of the optimal period, do show improvement in perceiving nonnative speech sounds after behavioral training for many languages, including Mandarin (e.g., Lively, Logan, & Pisoni, 1993; Pisoni, Aslin, Perey, & Hennesy, 1982; Wang, Jongman, & Sereno, 2003; Wang, Spence, Jongman, & Sereno, 1999). In fact, these phonetic training studies have highlighted the ability for the adult brain to remain plastic past puberty. Neuroimaging studies have demonstrated that with increased Mandarin proficiency, there is an increasing native-like left-lateralization pattern, characterized by increased cortical activation of preexisting language areas and the recruitment of additional cortical regions (Wang, Behne, Jongman, & Sereno, 2004; Wang, Sereno, Jongman, & Hirsch, 2003).

For native English speakers, some of the phonological difficulty is due to native English speakers tending to put more perceptual weight on pitch height rather than on contour (Wang, Jongman, & Sereno, 2006). For example, English speakers perceptually confuse the second and third tones with one another because they start at around the same F0 height, although the turning point occurs at different times, which is the crucial cue to differentiating between the two tones for native speakers. English speakers also make two kinds of tonal production errors—contour and tonal register—and the overall range is much smaller (Wang et al., 2006). As a result, they also do not distinguish between the two tones in production, dipping far too late for the second tone and, thus, forcing the falling and rising pitch range to be comparable to the third tone.

Nonetheless, many studies have shown that with perceptual and production training, Mandarin tone productions have improved in L2 learners (e.g., Wang, Jongman, et al., 2003). However, previous studies have mainly focused on how the aural/oral domain can aid in perceptual or production improvement (e.g., Wang et al., 1999; Wang, Jongman, et al., 2003). Recently, there has been a growing interest in exploring the intersection between this auditory phoneme learning and multimodal input, particularly in the perceptual realm (e.g., C. M. Chen, 2013; Hannah et al., 2017; Hirata & Kelly, 2010; Liu et al., 2011; Morett & Chang, 2015). We extend these Mandarin pitch discrimination studies into the production realm by investigating how producing metaphoric bodily actions can influence L2 tonal production from both an acoustic and perceptual perspective, in addition to exploring whether this influence of multimodal pitch information can be extended to L1 speakers as well.

The Multimodal Nature of Language Processing and Production

Lakoff and Johnson (1980) have argued for an embodied approach to language processing and learning. For example, Bolinger (1983) asserts that high and low pitches are the vocal representations of an up–down metaphor (see also Casasanto, Phillips, & Boroditsky, 2003); therefore, we can understand high and low pitches through the physical experience of the high/low muscle tension in rising and falling hand Gs and HNs, which parallels the high/low tension of the vocal tract muscles necessary to produce high or low pitches. This, together with theories that multiple streams of information facilitate information processing (Paivio, 1986), suggests that tonal production may be optimally viewed as a multimodal phenomenon.

Hand Gs. On the basis of theories that hand Gs comprise a fundamentally integrated system with speech during language production (McNeill, 2005), we are interested in the interplay between metaphoric hand Gs and the acoustics of speech production of Mandarin tones. There is a temporal synchrony to hand Gs and speech that makes verbal communication a deliberate act of both gesturing and speaking (McNeill, 2005), and this may have significant implications for L2 phoneme learning.

Gentilucci, Campione, Dalla Volta, and Bernardis (2009) found that both the observation and the execution of hand grasping toward differently sized objects influenced various aspects of an L1 speech. For example, when seeing a G reaching for a smaller object as opposed to a larger one, participants produced a syllable *da* with a narrower mouth aperture and the decreased first formant values. McClave (1998) found that the direction of intonation within English syllables was correlated with the direction of hand

Gs, that is, the rising and falling of the pitch tended to parallel that of the speakers' rising and falling hand movements, although this correlation is not "biologically mandated." Krahmer and Swerts (2007) found that native Dutch participants' manual beat Gs and HNs had significant effects on duration and the higher formants (e.g., the second formant) of the speech segment, whereas they did not necessarily affect the F0. Recent studies, however, have given more evidence that visuospatial information influences pitch perception, suggesting that auditory aspects of pitch and loudness share a spatial representation (Casasanto et al, 2003; Connell, Cai, & Holler, 2013; Lemaitre et al., 2017; Liu et al., 2011).

Regarding G and foreign language phoneme perception, the first studies to investigate this issue were training studies of Japanese (Hirata & Kelly, 2010; Hirata, Kelly, Huang, & Manansala, 2014; Kelly, Hirata, Manansala, & Huang, 2014). Rather than focusing on the role of G in phoneme processing across words in a sentence (cf. Krahmer & Swerts, 2007), these studies focused on how G might help people learn to hear novel phoneme contrasts within words in a sentence. These studies showed that training people to view or imitate metaphoric Gs, such as long and short sweeping hand Gs to illustrate the length of Japanese long and short vowels, did not assist native English speakers in learning to hear vowel length distinctions any better than training with speech alone. The conclusion was that, at this lowest level of processing individual foreign language phonemes, that is, differentiating phonemes within words, Gs are less integrated with speech than they are at higher levels of language comprehension (Kelly, 2017).

In contrast, research on Mandarin tone learning has shown that hand Gs do affect how English speakers perceive phonetic contrasts, or at least does not provide a cognitive burden, as was the case in Hirata and Kelly (2010). For example, Eng, Hannah, Leong, and Wang (2013) found that L2 speakers' tone identification in a four-alternative forced-choice task showed statistically significant improvement after a training condition seeing hand Gs trace the tonal contours while simultaneously viewing facial movements and listening to the speaker's audio. However, the amount of improvement before and after training in this auditory–visual–G condition was not significantly different from an auditory–visual condition or an auditory–G condition. Furthermore, Morett and Chang (2015) found that Gs that mimicked Mandarin tonal contours enhanced native English speakers' identification of the meaning of words that were minimal lexical tone contrasts. However, subjects' auditory ability to identify the Mandarin tones improved only in the same amount as an audio-only condition. In contrast, very recent research has shown that viewing hand Gs along with facial articulatory cues significantly improved native English speakers' identification of Mandarin tones in a noisy background over an auditory–facial articulatory cues condition alone (Hannah et al., 2017; see also Kelly, Bailey, & Hirata, 2017, for a similar result focusing on Japanese sentence-final syllable contrasts). Building on this research in perception, we hypothesize that manual Gs may also play a role in Mandarin tonal production.

HNs. Although HNs are not explicitly included in McNeill's theory of G–speech integration, we still have reason to believe that these head movements are tightly tied to speech comprehension and production (see Loehr, 2007). HNs are naturally produced for prosodic purposes in languages, such as Mandarin and Japanese, simultaneously with speech (T. H. Chen & Massaro, 2008; Munhall, Jones, Callan, Kuratate, & Vatikiotis-Bateson, 2004). Therefore, if speech perception is regulated by the motor system, HNs may play a role in language comprehension. Indeed, Munhall et al. (2004) found that HNs that mirrored the prosody of Japanese allowed for improved L1 speech comprehension in noisy environments. This phenomenon is found not only for prosody but also for lexical tones: Burnham, Lau, Tam, and Schoknecht (2001) and Burnham et al. (2006) found that both native and nonnative perceptions of Cantonese lexical tones relied heavily on rigid head motion as a perceptual cue. This finding extends to the identification of Mandarin lexical tones: T. H. Chen and Massaro (2008) demonstrated that L1 speakers of Mandarin could identify the tones significantly above chance without sound just by relying on visual head movement cues alone.

This Study

Language comprehension and learning actively recruit multimodal information—from the mouth, face, hand, and head—in both the L1 and L2 domains (e.g., Burnham et al., 2001, 2006; Chen & Massaro, 2008; Kelly, 2017; Munhall et al., 2004). Similarly, language production relies on Gs and facial expressions to convey additional information to form an integrated G–speech system (e.g., Chen & Massaro, 2008; Gluhareva & Prieto, 2017; Krahmer & Swerts, 2007; Loehr, 2007; McClave, 1998). Given the multimodal nature of language, people should take advantage of this multimodal aspect rather than relying solely on the aural/oral domain with no bodily input.

Much of the literature on HNs and hand Gs has dealt with L1 speakers but not with L2 speakers. Of the studies with L2 speakers, the focus has been on perceptual discrimination rather than on production. If there is a neuromotor link between speech and Gs, it is important to determine the extent to which these two are linked and on what levels they operate. There is a burgeoning evidence that, perhaps, this speech–G integration breaks down at the phonological level with respect to phonemic length contrasts (e.g., Hirata & Kelly, 2010); however, on the other hand, speech and G may still be very well integrated when concerned with pitch perception (Eng et al., 2013; Hannah et al., 2017; Kelly, 2017; Morett & Chang, 2015). Therefore, there is a need to investigate whether L2 native English speakers and L1 native Mandarin speakers—two groups at both ends of the Mandarin fluency spectrum—can also recruit these head movements and manual Gs. Broadly, we aim to define the boundaries and limitations of bodily actions (HNs and hand Gs) on tonal production.

In Experiment 1, native Mandarin speakers produced Mandarin words under three conditions that varied on the type of information presented: speech only (SO) control, speech + Gs, and speech + HNs. Experiment 2a followed the same protocol as Experiment 1; however, the participants were monolingual native English speakers with no experience speaking Mandarin. In Experiment 2b, a new set of native Mandarin speakers judged the overall accuracy of native English speakers' tonal productions from Experiment 2a.

Experiments 1 and 2 together form a comprehensive picture of how Gs and HNs might influence tonal production from both ends of the fluency spectrum. On the one end, native Mandarin speakers have a mature Mandarin phonological system that is established enough to sustain an integrated motor system for speech. On the other end, we have native English speakers whose Mandarin phonological system is very immature. Using these two ends of the continuum, we examined whether hand and head movements affect this low-level phonemic processing, specifically in the production domain. We made three competing predictions: (a) Based on research showing that Gs do not play a role in learning low-level phonemic length perceptual discrimination in Japanese (Hirata et al., 2014; Hirata & Kelly, 2010; Kelly et al., 2014), metaphoric pitch Gs should not affect the production of Mandarin lexical tones; (b) based on work showing that metaphoric *pitch* Gs do play a role in perceiving Japanese (Kelly, Bailey, & Hirata, 2017) and Mandarin pitch distinctions (Eng et al., 2013; Morett & Chang, 2015), metaphoric pitch Gs should affect the production of Mandarin lexical tones; and (c) if the proficiency level modulates bodily influence on tone production, we may see an effect of hand Gs and/or HNs on Mandarin tonal contrasts in one group but not in the other group. Our study will be one of the first to systematically explore the role of metaphoric bodily actions in helping native English speakers to articulate—rather than to simply auditorily discriminate between—tonal contrasts.

Experiment 1: L1 Speakers

Participants

Twelve female native Mandarin speakers (18–22 years old), who were university students in the United States, participated. They had no known hearing issues, tone deafness, or previous music training.

Method

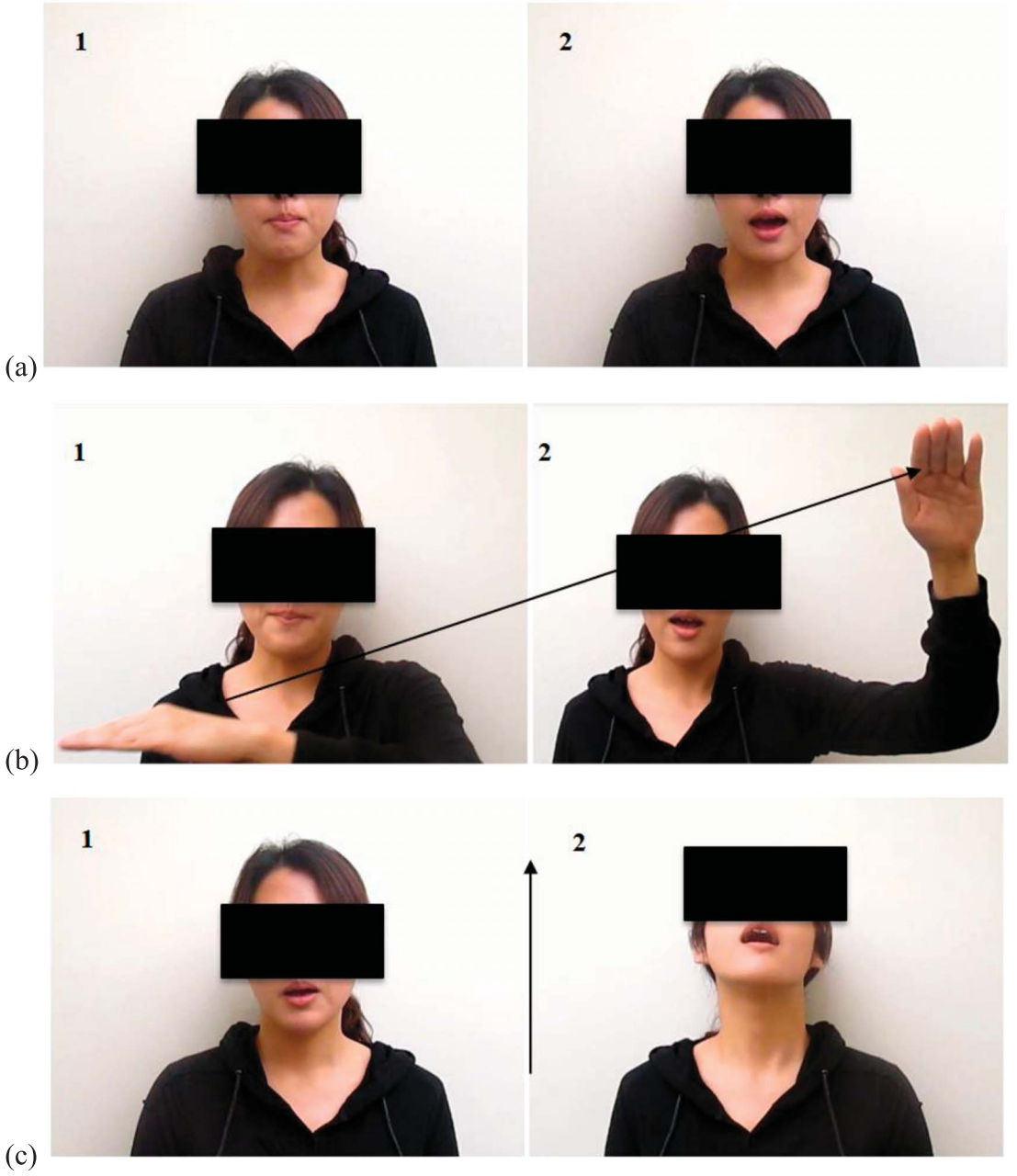
Stimuli. The experiment was administered to both L1 and L2 speakers with the same stimuli and design; however, it was created with L2 speakers in mind. Thus, the descriptions below make more sense in light of Experiment 2. Tones were produced by a native female speaker from Beijing who speaks standard Mandarin Chinese. We recorded her audio separately and then filmed her lip-synching to her own audio under the three conditions. The native model recorded 12 different words for the experiment: three syllables ("ma," "mi," and "mu") × four tones. Two words of the same syllable, e.g., ma1–ma4, were presented in pair as one trial. These are real words except that the "mu" with the first tone is a nonsense word. The recorded audio was then dubbed over the edited videos to ensure uniformity of model audio. All the audio clips from the model were about half a second in length regardless of syllable. The stimuli were produced by only one speaker to reduce mental fatigue for our Mandarin-naive subjects in Experiment 2. Although there is research to suggest that speaker variability is important in enhancing the learning of nonnative speech sounds, especially in generalizing to novel stimuli (e.g., Lively et al., 1993), we chose to have single-speaker conditions because the principal focus was to observe the effects of hand Gs and HNs on the accuracy of tonal production and not on learning as defined by generalization to novel target sounds.

Monosyllabic words were chosen instead of whole phrases or sentences because developmental accounts of L2 tone production by English-speaking students show that they go through two phases of tone acquisition. Students learn to produce the canonical tones in rehearsed speech before gaining mastery over tone production in a sentence-long or paragraph-long spontaneous speech where the tonal contours are not fully produced (Tao & Guo, 2008). Because our L2 participants were unfamiliar with Mandarin, we wanted to expose them to isolated monosyllabic words. Furthermore, because the primary focus was on lexical tone production alone, we chose Mandarin syllables that could also be found in English ("ma" /ma/, "mi" /mi/, and "mu" /mu/), avoiding unfamiliar nonnative segments that would be difficult to learn.

The tones were presented in three different within-subject conditions: (a) no hand Gs/HNs, which act as the SO control, (b) speech + hand Gs, and (c) speech + HNs. Gs and HNs were presented simultaneously with the native model's audio, which was acoustically identical across all three conditions for each syllable. The hand Gs and HNs metaphorically reflected both the height and the contour of the tones intentionally. The video of the native Mandarin speaker model framed her so that she was in the center of the frame from the chest up. In the hand Gs and HNs videos, the model made use of the entire range of the frame. In all videos, the model also made sure to enunciate very clearly. Thus, visual information from both the vocal articulation and the Gs/HNs should have been very salient to the participants. In the SO condition (see Figure 1a), the videos present the native model only speaking without moving her head or her arms in any way.

In the hand Gs condition (see Figure 1b), the native model speaks while simultaneously moving her hand from left to right to metaphorically reflect both the contours and the relative heights of the tones. A high pitch was in the range of the top of the head, a middling pitch at above shoulder height, and a low pitch at the chest level. Therefore, the first tone has the hand moving from left to right in a straight line at the height of the top of her head. For the second tone, the hand starts at the shoulder height and moves upright diagonally to end at the height of the top of

Figure 1. The model is presenting the second tone (ma2) in all three panels. In (a), control condition, she says it with speech without moving the head and hands. In (b), gesture condition, the model moves her hand from her left shoulder across her body to the right side of her head, indicating a midrising tone. In (c), head nod condition, the model starts with her head looking forward (indicating a middling pitch) and, then, moves it straight upward to represent a rising pitch.



her head. The third tone has the hand start at the shoulder height, dip to the chest level and, then, rise again to the shoulder height. The fourth tone starts the hand at the top of the head and drops down to the chest level.

With regard to the HNs (see Figure 1c), a high pitch is encoded with an upward tilt, a middling pitch has the head at a normal straight-on position, and a low pitch positions the head so that it is tilted downward toward the chest. Therefore, the first tone has the head kept in the upward tilt for the entire duration. The second tone moves the head from the straight-on level position to an upward tilt. The third tone dips the head from the straight-on level position to the chest and, then, back to the level position again. The fourth tone starts with the head at an upward tilt and, then, forces the head to dip toward the chest. As a result, we see that both the hand Gs and the HNs encode

a comparable amount of information about the Mandarin tones in terms of the tonal contour and the height.

The order of the presentation of the multimodal conditions was counterbalanced. Within each of the three condition blocks, there were 15 trials—each consisting of a 2-s-long video that presents two monosyllabic words (e.g., ma1 and ma2), followed by a "3-2-1" countdown and, then, a "please repeat" cue. Participants then had to reproduce the two words from memory (see Procedure subsection). Having pairs instead of single words within a trial was to maintain participants' interest while also enabling them to hear tonal differences.

Procedure. Participants were introduced to the Mandarin tones and conditions, using a monosyllable "yi" (/*il*/), which was not a syllable in the actual experimental trials. There was first a familiarization video in which the experimenter explicitly highlighted the acoustic differences among the four tones, in addition to demonstrating how the Gs were mapped onto the spoken tones. Participants played the videos multiple times until they understood what the four tones were and what the conditions would entail. To quickly check understanding, the researchers quizzed the participants, "Which tone is X?" or "If the model goes like this [speak and demonstrate HN/hand G], what will you do?" using the dummy syllable "yi." Because the research paradigm was not for perceptual discrimination or identification, participants did not undergo a more vigorous initial perceptual task. Rather, participants were told to imitate exactly what they have seen and heard in all aspects to the best of their abilities.[1] Participants were audio-recorded using the Computerized Speech Lab model 4150 machine with a 22050-Hz sampling rate. Researchers observed the participants to ensure task compliance, instructing participants to speak loudly and produce Gs and HNs confidently.

Analysis. On the basis of previous research on tone production (e.g., Wang, Jongman, et al., 2003), we focused on F0 acoustic analyses. These have been the focus of past works primarily because they are the fundamental perceptual cues for tones (e.g., Chang, 2011; Lieberman & Blumstein, 1988; Wang, Jongman, et al., 2003; Wang et al., 2006). An acoustic analysis of the F0 values using Praat was undertaken to determine any conditional effects of manual Gs or HNs on Mandarin tonal production (Boersma & Weenink, 2015). F0 values were sampled at every quintile of trial duration, that is, 0% (onset), 25%, 50%, 75%, and 100% (offset) of the way through each speech segment (based on Wang, Jongman, et al., 2003). All statistical analyses were conducted in R (R Core Team, 2016). Using the lme4 package, we ran a linear mixed-effects model using F0 values as the response variable with condition, tone, and time as fixed effects—and trial nested within subjects and word as random effects—to examine conditional effects on the tonal contour for both L2 and L1 speakers in Experiments 1 and 2 (Bates, Maechler, Bolker, & Walker, 2015). The models included varying intercepts for the participant to account for the F0 variation in participants' voices and trial to reflect the repeated-measures experimental design and to account for item effects. Significance tests for predictors were performed using the car library in R with Type II Wald chi-squared tests (Fox & Weisberg, 2011). Because we are primarily concerned with the effect of hand Gs or HNs on F0 values, which vary across time, we focused on the three-way Condition × Tone × Time interactions, indicating that there were conditional effects on F0 values across every time point for a tone. We used the two-way Tone × Time interaction effect as a stimulus check because F0 values should vary across time for each tone as a function of a tone's F0 contour. Estimated means, standard errors, and post hoc pairwise comparisons for significant interaction effects were calculated with R's lsmeans package (Lenth, 2016). All p values of post hoc contrasts reported were adjusted for multiplicity with the mean value theorem (mvt) method for each tone (Lenth, 2016). Post hoc comparisons allowed us to examine conditional differences within the same time point of a tone, but also it also allowed us to explore F0 "contour changes" within a tone of each condition on the basis of the expected shape of a tone (see Results of Experiment 2a for more details). The same procedures were followed for an analysis in Experiment 2.

Results

After running a chi-square test of significance on the L1 speaker model, we found a significant main effect of condition, $\chi^2(2) = 6.20$, $p = .036$. However, this effect is difficult to interpret considering what an "averaged tone" at an "average time point" actually means. More importantly, we did not find a significant three-way Condition × Time × Tone Effect interaction, $\chi^2(24) = 20.25$, $p = .683$, (see Table 1).

Table 1. The chi-square test of significance on the predictors of the L1 speaker model in Experiment 1 did not yield a significant three-way Condition × Time × Tone interaction but did a two-way Time × Tone interaction.

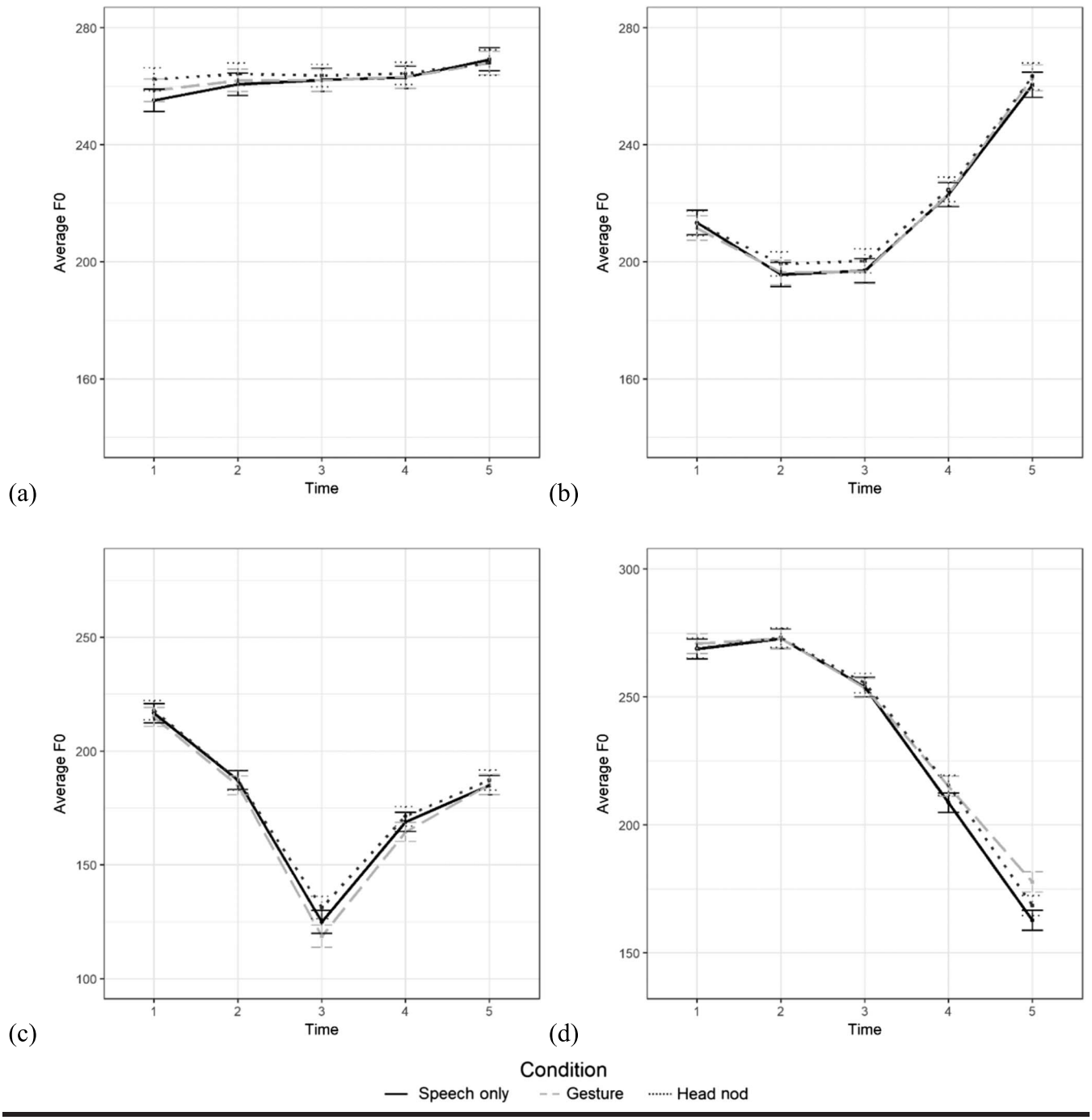
L1 speaker model: predictors' test of significance			
Factor	χ^2	***df***	**Pr(> χ^2)**
Condition	6.20	2	0.036*
Time	1058.70	4	< 0.001***
Tone	4775.70	3	< 0.001***
Condition × Tone	7.11	6	0.311
Condition × Time	9.16	8	0.329
Time × Tone	6522.95	12	< 0.001***
Condition × Tone × Time	20.25	24	0.683

*$p < .05$; ***$p < .001$.

[1]Note that our task combined the observation and production of HNs and hand Gs. This conflation makes it impossible to pull apart which part—observing or producing—has an effect on the acoustics of the Mandarin tones. However, this situation most accurately captures what happens in actual introductory Mandarin classrooms, where instructors commonly produce actions, such as HNs and hand Gs, to illustrate tonal differences and, then, ask students to mirror those actions themselves when repeating the tones.

This suggests that there were no conditional differences in native Mandarin speakers between Gs, HN, and SO conditions for all four tones at all five time points at which F0 values were sampled (see Figures 2a–d). Because there was not a significant three-way interaction, we did not run post hoc contrasts to examine conditional differences within the same Tone × Time combination nor conditional differences in F0 changes across time within a tone. It is worth pointing out that this lack of an effect was not due to an underpowered design; for example, we did clearly see the expected two-way Tone × Time interaction, $\chi^2(12) = 6522.95$, $p < .001$, revealing that F0 values for each tone

Figure 2. The L1 speakers' estimated F0 values and standard errors for the first (a), second (b), third (c), and fourth (d) tones in Experiment 1. There are no differences in condition at any time point for all four tones. Average F0 were in Hz, and time points 1–5 represented each quintile of the speech segment.



varied significantly as a function of time, which one would expect due to tonal contours.

Thus, at least for L1 speakers, producing hand Gs and HNs with speech does not reliably affect the F0 when producing Mandarin tones.

Experiment 2: L2 Speakers

Experiment 2 had two phases. In the first phase (Experiment 2a), we conducted the same paradigm as in Experiment 1 with nonnative Mandarin speakers, in which the F0 of their production was analyzed. In the second phase (Experiment 2b), we used a group of native Mandarin speakers as judges to rate the perceptual accuracy of the L2 speakers' production. Experiment 2b provided an additional measure besides the F0 analysis to examine whether their production significantly differed across three conditions. The F0-dependent measure might reveal only minor differences that may not be properly "perceived" by native judges in terms of the four lexical tone categories, or the native speaker perception might capture some conditional differences that were not reflected on the F0 measure alone (e.g., syllable duration), or, in the case that the F0 measure did reveal conditional differences, the judges' scores might provide another point of converging evidence that metaphoric hand Gs and/or HNs did help in L2 tonal production. In other words, we asked, "Can the L2 articulations pass the ear test?"

Participants

Experiment 2a. In the first phase, 24 female monolingual native English speakers (aged 18–22 years) were recruited. All the participants were screened with a linguistic background survey to ensure that they had minimal foreign language experience and no knowledge of Mandarin. No participants had any known hearing problems or any music training. The participants were all female to reduce the variability in pitch and facilitate acoustic analysis.

Experiment 2b. In the second phase, seven native Mandarin speakers acted as judges to determine which tone they thought that the L2 speakers in Experiment 2a had produced for each trial. Before a native speaker could be enrolled as a judge, we ensured that they were competent at categorizing Mandarin tones. They were administered a language survey to confirm that their native language was Mandarin and that they both spoke and understood the standard variety of Mandarin. As a final step of confirmation, judges were presented with the native model stimuli from Experiment 1 and were asked to identify the tones they heard—all judges were near ceiling in this task. Thus, all seven raters can be considered reliable speakers of Mandarin.

Method

Stimuli and procedure. The same stimuli and procedures used in Experiment 1 were used in Experiment 2a. However, an additional questionnaire was added to the end of Experiment 2a to assess participants' opinions on the utility of the metaphoric bodily actions. Specifically, at the end of the task, each L2 speaker was asked if he or she found the Gs and HNs helpful. They were also asked to rank the three conditions in order of helpfulness.

In the second phase, L2 speaker productions from Experiment 2a were used as stimuli for native judges to identify which of the four tones they perceived. Each block consisted of 15 trials spoken by an L2 participant for each condition; and the order of the blocks was randomized so that native judges would not become normalized to any single L2 speaker. Each trial consisted of an L2 speaker's tonal "minimal pair" from Experiment 2a followed by 2.5 s to allow for judges' responses. Participants then identified each L2-produced word as the first, second, third, or fourth tone, or none if it was too ambiguous to be judged. Responses were marked on a sheet with all the tones of /ma/, /mi/, and /mu/ (except for the nonsense word /mu1/) written out in Chinese characters at the top of the sheet as a reference (e.g., 妈, 麻, 马, 骂 for /ma1, 2, 3, 4/). Judges listened to all 72 subject–condition blocks (24 L2 speakers × three conditions) for a total of 1,080 trials. Audio output was produced through a speaker right in front of the judges.

Analysis. In Experiment 2a, the same within-subject analysis applied to the L1 speaker model was applied to the L2 speaker model (as detailed in the previous subsection; see F0 Scores). We ran a Type II Wald chi-square test of significance on the fixed effects of the L2 speaker model. For the post hoc contrasts for significant predictors, not only did we run analyses to compare conditional differences between the same time points within a tone but we also looked at F0 changes between time points to examine conditional differences on the tonal contour itself. For the self-report data, binomial tests were conducted to ascertain if the proportion of individuals who found Gs and HNs helpful exceeded the expected probability (see Self-Report Scores subsection).

In Experiment 2b, a trial was defined as being *correct* if all seven judges agreed and *incorrect* if the decisions were not unanimous. Accuracy was then calculated by taking the percentage of correct trials per condition per subject. We then examined if accuracy differed across conditions using a linear mixed-effects model with accuracy as the response variable, condition and tone as the predictors, and subject as a random effect (see Native Judge Ratings subsection). On the basis of the data collected from the native Mandarin-speaking judges, we also reanalyzed the L2 speakers' F0 values using only correct trials, that is, trials that all seven judges unanimously agreed were correct. We reran the same L2 speaker regression model on this data subset (see Reanalyzing L2 Speakers (Experiment 2a): Correct-Only Trials subsection). Post hoc contrasts for the model reanalyzed the shape of the F0 contours (i.e., F0 differences between two time points) and examined their conditional differences.

Results: Experiment 2a

F0 scores. We found the expected significant two-way Time × Tone interaction, reflecting the unique F0 contour of each tone, $\chi^2(12) = 2739.91$, $p < .001$, and showing that L2 speakers were clearly able to produce differentiable tones. There were also conditional differences as evidenced by a significant three-way Condition × Tone × Time interaction, $\chi^2(24) = 55.21$, $p < .001$. We ran post hoc pairwise comparisons on this three-way interaction to examine (a) the conditional differences within the same tone and time point and (b) the conditional differences of F0 differences across time within a tone. The latter post hoc analysis allowed us to capture the effect of condition on an F0 range (Table 2), motivated by the shape of each tone's respective pitch contour.

Figures 3a–3d demonstrate the mean estimated F0 values and their standard error for each tone for all five time points at each condition (see Figure 3). Post hoc contrasts revealed that, at speech offset (Time Point 5) for the fourth tone, the F0 value in the G condition was significantly lower than that of the HN or SO conditions ($z = -3.06$, $p = .03$; $z = -3.14$, $p = .02$, respectively). At all other time points for the other three tones, there were no significant differences in condition.

Because the first tone is characterized by its high, steady pitch, we looked at the conditional differences of the F0 differences between Time Point 1 and 5. The post hoc analysis of the three-way Condition × Tone × Time interaction revealed no significant conditional differences between SO and G (SO = −6.3 Hz, G = −2.86 Hz, $z = -1.01$, $p = .57$), SO and HN (SO = −6.63 Hz, HN = −3.15 Hz, $z = -0.94$, $p = .62$), and G and HN (G = −2.86 Hz, HN = −3.15 Hz, $z = 0.08$, $p = .99$).

The second tone is characterized by a small dip in the beginning with the turning point around Time Point 2 before rising sharply again. We looked at the F0 dip between Time Points 1 and 2 and the F0 rise between Time Points 2 and 5 for the effects of condition. There were no significant differences in condition in the F0 dip between SO and G (SO = −14.43 Hz, G = −12.63, $z = -0.40$, $p = .99$), SO and HN (SO = −14.43 Hz, HN = −15.88 Hz, $z = 0.32$, $p = .99$), and G and HN (G = −12.63, HN = −15.88 Hz, $z = 0.71$, $p = .93$). There were no significant differences in the F0 rise between SO and G (SO = 44.88 Hz, G = 51.55 Hz, $z = -1.47$, $p = .48$), SO and HN (SO = 44.88 Hz, HN = 46.86 Hz, $z = -0.43$, $p = .99$), and G and HN (G = 51.55 Hz, HN = 46.86 Hz, $z = 1.02$, $p = .78$).

Similarly, the third tone is also characterized by a dip and a rise with the turning point around Time Point 3. However, we also did not find any significant conditional differences in the F0 dip between SO and G (SO = −25.62 Hz, G = −26.44 Hz, $z = 0.18$, $p = .99$), SO and HNs (SO = −25.62 Hz, HN = −37.03, $z = 2.50$, $p = .06$), and G and HNs (G = −26.44 Hz, HN = −37.03, $z = 2.34$, $p = .92$). We did not find any conditional differences in the F0 rise either between SO and G (SO = 48.67 Hz, G = 52.68 Hz, $z = -0.88$, $p = .86$), SO and HN (SO = 48.67 Hz, HN = 55.77 Hz, $z = 01.56$, $p = .43$), and G and HN (G = 52.68 Hz, HN = 55.77 Hz, $z = -0.68$, $p = .94$).

As the fourth tone's characteristic is a sharp fall in F0, we took the difference in the F0 between Time Points 1 and 5 to investigate differences in condition. There was a significant difference in the F0 fall between G and SO conditions (G = −51.52 Hz, SO = −38.28 Hz, $z = 3.72$, $p < .001$) and between G and HN (G = −51.52 Hz, HN = −42.77 Hz, $z = -2.37$, $p = .05$) but not between SO and HN (SO = −38.28 Hz, HN = −42.77 Hz, $z = 1.34$, $p = .37$).

As a quick summary, there were no conditional differences for the first, second, and third tones, but there was a significant effect of hand G for the fourth tone: With the hand G, the nonnative speakers' F0 went down at the end of the syllable more than that of the HN and SO conditions.

Self-report scores. Self-report data from the L2 speakers indicated that 19 out of the 24 participants found that the G condition helped them pronounce the tones correctly. A binomial test indicated that this was greater than the expected proportion (50%, $p < .001$). Seventeen participants also found the HN condition similarly helpful, which was not significantly above expectations (50%, $p = .06$). When ranking the conditions in order of helpfulness, 19 participants chose the G condition as their number one ranking; two chose the HN condition, and three chose the SO condition. The probability of ranking G as most helpful was significantly above chance (33.33%, $p < .001$).

Table 2. The chi-square test of significance on the predictors of the L2 speaker model in Experiment 2a found significant three-way Condition × Time × Tone interaction and two-way Time × Tone interaction.

L2 speaker model: predictors' test of significance

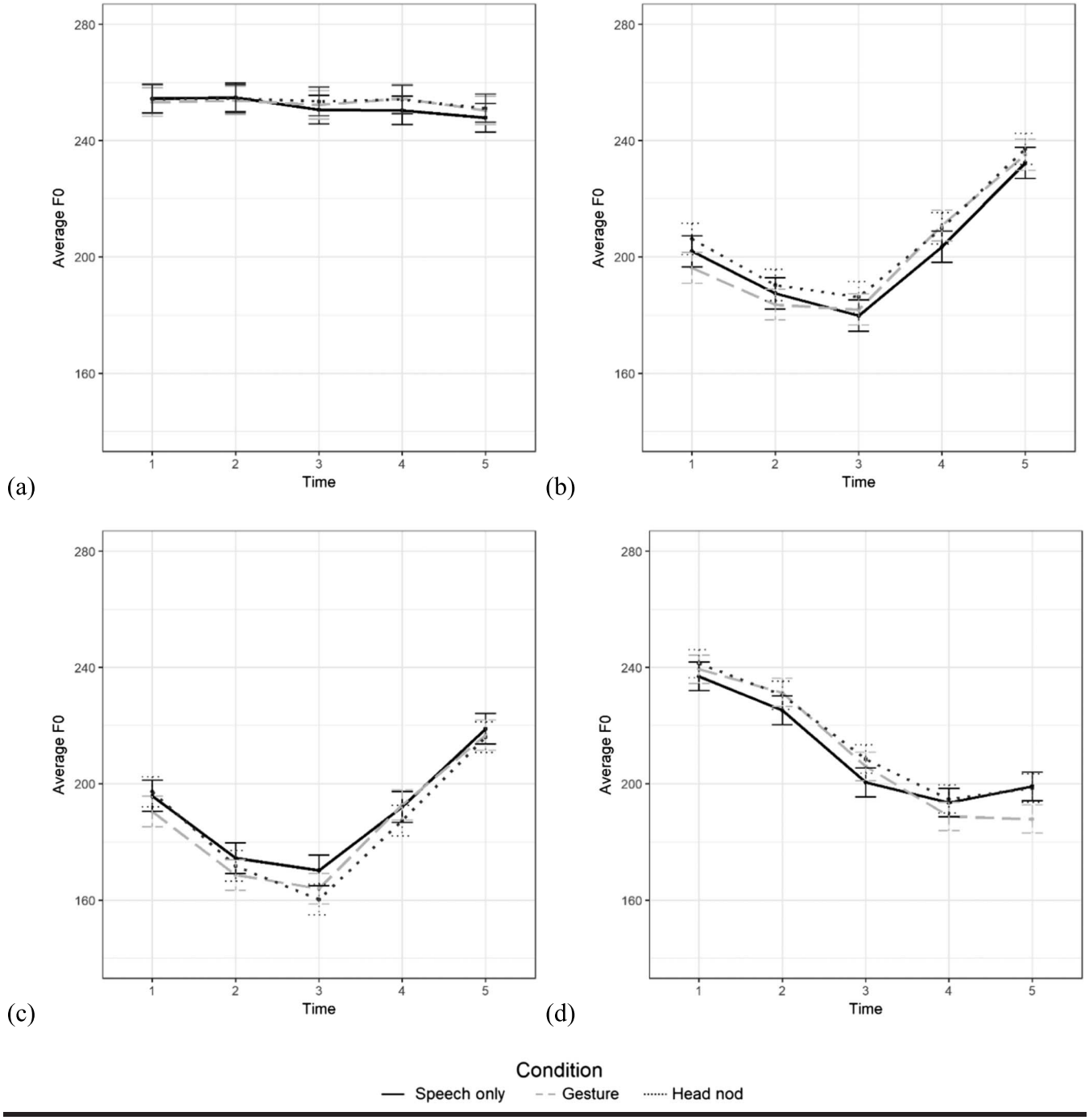
Factor	χ^2	*df*	Pr(> χ^2)
Condition	2.77	2	0.250
Time	783.58	4	< 0.001***
Tone	274.63	3	< 0.001***
Condition × Tone	5.09	6	0.532
Condition × Time	8.19	8	0.416
Time × Tone	2757.61	12	< 0.001***
Condition × Tone × Time	52.57	24	0.001***

***$p \leq .001$.

Results: Experiment 2b

Native judge ratings. There was not a main effect of condition, $\chi^2(2) = 1.76$, $p = .42$, nor a two-way Condition × Tone interaction, $\chi^2(6) = 2.03$, $p = .92$, suggesting that condition did not influence L2 speakers' accuracy rates. However, there was a main effect of tone, clearly demonstrating that participants struggled more with certain tones than others, $\chi^2(3) = 284.26$, $p < .001$ (see Table 3). Based on the estimated accuracy rates from the regression model, participants found the first tone to be the easiest to reproduce (81.65%), followed by the fourth tone (47.82%), third tone (38.00%) and, then, the second tone (16.97%). A post hoc analysis

Figure 3. The L2 speakers' estimated F0 values and standard errors for the first (a), second (b), third (c), and fourth (d) tones in Experiment 2a. Average F0 were in Hz and time points 1–5 represented each quintile of the speech segment. In (d) at speech offset, we see conditional differences such that the gesture condition (F0 mean = 187.854 Hz) is significantly different from both the head nod (F0 mean = 198.571 Hz) and speech-only (F0 mean = 198.881 Hz) conditions by an estimated −10.716 Hz (p = .029) and −11.027 Hz (p = .023), respectively.



revealed that the accuracy rates for the first tone were significantly different from that of the fourth tone (Δ = 33.83%, z = 8.45, p < .001), although accuracy rates for the fourth and third tones were similar (Δ = 9.82%, z = 2.45, p = .07). Accuracy rates for the second tone fell significantly below that of the third tone (Δ = 21.03%, z = 5.26, p < .001; see Figure 4).

Reanalyzing L2 speakers (Experiment 2a): Correct-only trials. There was still a significant three-way Condition × Tone × Time interaction, $\chi^2(24) = 42.63$, p = .01 (see

Table 3. The chi-square test of significance on the predictors of the accuracy model from Experiment 2b revealed a significant main effect of tone but not a two-way Condition × Tone interaction nor a main effect of condition, suggesting that there were no conditional effects on accuracy.

L2 speaker accuracy model: predictors' test of significance

Factor	χ^2	*df*	Pr(> χ^2)
Condition	1.76	2	0.416
Tone	284.26	3	< 0.001***
Condition × Tone	2.03	6	0.917

***$p < .001$.

Table 4. The chi-square test of significance on the predictors of the reanalyzed L2 speaker model from Experiment 2a with a subset of trials consisting of correct-only trials that were determined in Experiment 2b.

Correct-only L2 speaker model: predictors' test of significance

Factor	χ^2	*df*	Pr(> χ^2)
Condition	1.97	2	0.37435
Time	893.16	4	< 0.001***
Tone	314.50	3	< 0.001***
Condition × Time	9.56	8	0.298
Condition × Tone	3.90	6	0.690
Time × Tone	3764.21	12	< 0.001***
Condition × Time × Tone	42.63	24	0.011*

Note. There is still a significant three-way Condition × Tone × Time interaction.

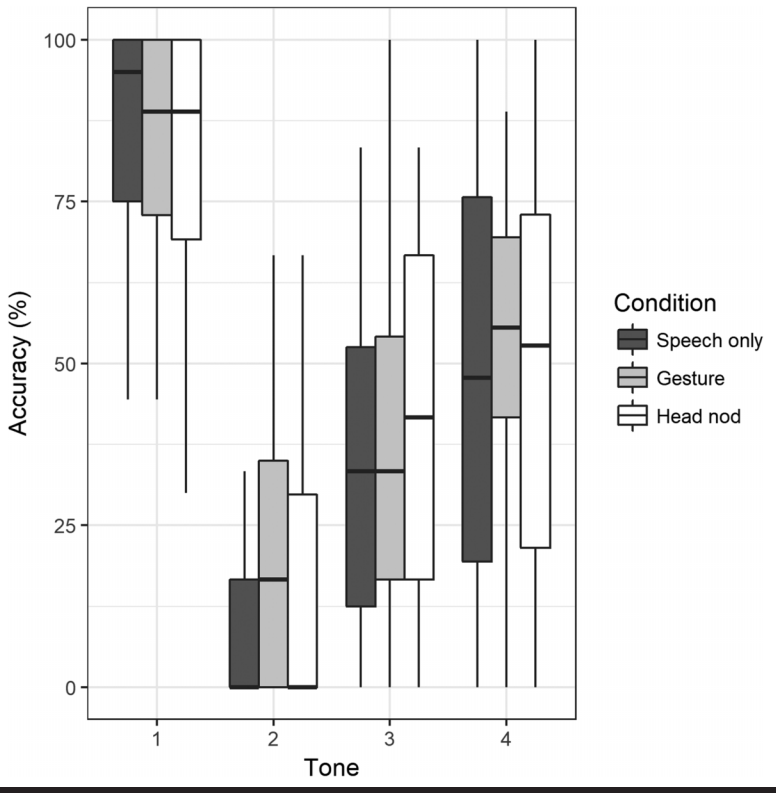
*$p < .05$; ***$p < .001$.

Table 4). However, post hoc contrasts with multiplicity adjustments for each tone revealed no significant differences of condition for all tones at all time points. Based on this significant three-way interaction, we also reanalyzed the shape of the pitch contours by investigating the conditional differences of F0 differences across time. For the first tone, we found no significant conditional differences on the F0 difference between speech onset (Time Point 1) and offset (Time Point 5). There was not a significant difference between conditions in the second tone's F0 dip and rise to/from the turning point at Time Point 2. However, there was an HN effect (−47.09 Hz) in the third tone's F0 dip (from speech onset to Time Point 3) compared with the SO condition (−27.10 Hz; $z = 3.21$, $p = .01$), such that HN dipped 19.99 Hz lower than SO. The fourth tone's F0 fall from speech onset to offset revealed no significant conditional differences either.

General Discussion

In Experiment 1 with native Mandarin speakers, hand Gs and HNs had no observable influence on the F0

Figure 4. Mandarin-speaking judges' correct identification of tones produced by L2 speakers (Experiment 2b). There are no differences in condition at any time point for all four tones.



values for all four tones. However, the results were somewhat mixed for nonnative speakers in Experiment 2: Although the vast majority of the findings were comparable to Experiment 1, there were hints that hand Gs influenced F0 productions for the fourth tone (Experiment 2a) and HNs affected F0 productions for the third tone (Experiment 2b). Taken together, the safest conclusion is that the metaphoric bodily actions of hand Gs and HNs have no effect on tonal productions for L1 speakers, and they may have only slight benefits for speakers with no knowledge of Mandarin. Thus, proficiency levels may modulate a G's influence on tone production, with a modest effect of hand Gs and HNs on Mandarin tonal contrasts in novices but not in experts.

Replication of Previous Studies

Before exploring the mixed role of HNs and hand Gs for L1 and L2 speakers, it is worth highlighting that we replicated the previous findings on L2 speakers' difficulties with Mandarin lexical tone acquisition (e.g., Jongman et al., 2006; Wang et al., 2006). Namely, our participants did not have a wide-enough F0 range compared with L1 speakers to correctly capture the tonal contours—basically, they did not dip their pitch low enough or raise it high enough. Furthermore, we saw that certain lexical tones were more difficult for L2 speakers to produce correctly. Consistent with previous research (e.g., Wang et al., 2006), L2 speakers had the easiest time with the first ("high, flat") tone; however, the second ("rising") tone had the lowest accuracy rate and was often confused with the third ("dipping") tone, as can be seen in similarity of the shapes of the estimated pitch contours (see Figures 3b and 3c). This replication is useful for framing the nonsignificant results regarding hand Gs and HNs: It suggests that our subjects took the task seriously and produced clean and reliable enough data, making it easily interpretable in the context of other studies on L2 Mandarin learning.

Roles of Hand Gs and HNs for L1 Speakers

In Experiment 1 with native Mandarin speakers, there was no evidence that hand Gs or HNs affected a standard measure (F0) of tonal production in L1 Mandarin speakers. This finding is interesting considering other research showing that bodily actions are tightly tied to L1 speech production (Krahmer & Swerts, 2007; Loehr, 2007; McClave, 1998; Munhall et al., 2004). For example, Loehr (2007) found that eye brow movements correlate with certain speech contours of native English-speakers, and Krahmer and Swerts (2007) showed that brow movements, together with HNs and manual beat Gs, changed the acoustic properties of the co-occurring speech (in duration and formants 2 and 3). Why the inconsistency? One major difference between this study and previous research is that we have focused specifically on the influence of bodily actions on the smallest of speech units: a single phoneme. This contrasts with the previous studies examining the effects of Gs and HNs on the span of multiple phonological units of a spoken utterance, for example, hand G directing attention to a word in a stretch of utterance. Indeed, Kelly (2017) has argued that bodily actions, such as hand G, are not naturally designed to play a role at the lowest levels of language processing and are much more optimally suited for higher levels (see also Hirata et al., 2014; Kelly et al., 2014).

It is important to note that this null finding should not necessarily be taken as evidence that hand Gs and HNs are poorly connected to speech in L1 tonal production. After all, it is possible that these bodily actions were shaped by an already mature L1 phonological system. That is, because native Mandarin speakers are experts at producing the four tones, it is possible that, when they were asked to also produce corresponding hand Gs and HNs, the bodily actions simply followed along, in a similar way a toddler is led by the hand of a parent. So it may still be the case that the L1 speakers move their heads and hands with their tones in a coupled way, but because the heads and hands are just following along, they do not influence the tonal contours. From this, we would predict that if we had video-recorded the Gs of our native speakers, they would be much more exaggerated than the Gs of our nonnative speakers. Alternatively, if we had instructed speakers to produce incongruent tone–G pairs, we might see effects of Gs on L1 production.

It is also important to note that the hand Gs we used in this study refer to only a type of metaphoric G that imitates the nature of the sounds themselves and not the meaning of words, such as iconic Gs (e.g., a G of hand holding a cup and imitating the action of drinking). These iconic Gs represent the meaning of words, and they likely have a qualitatively different relationship with the speech they accompany, having significant semantic and pragmatic influences on both L1 and L2 processing (Kelly, 2017).

Roles of Hand Gs and HNs for L2 Speakers

The story is slightly more complicated with L2 speakers. For the most part, Experiment 2a replicated the results of Experiment 1, showing that there was again no significant F0 difference among our conditions for all tones and time points, with the lone exception of a speech offset for the fourth tone (the "falling" tone) in the G condition. Even when examining the F0 differences between time points within each tone to characterize the tonal contour shape, there were no conditional differences except for the final time point in Tone 4, with the G condition dropping slightly lower than the two other conditions. It should be noted that this conditional difference was found across all speech samples, regardless of the perceived accuracy by native speaker judges in Experiment 2b. In other words, an exaggerated falling hand G facilitated the production of a falling pitch, but that did not seem to facilitate its perception as a canonical Tone 4.

In Experiment 2b, native Mandarin speakers acted as judges to assess the accuracy of L2 speakers' tonal productions from Experiment 2a. We found that the condition did not affect the accuracy rates of L2 speakers' tonal

productions from an overall perceptual perspective. This conclusion seemed to hold even when conducting an F0 analysis on only the subset of unanimously "correct" trials for the L2 speakers, with no significant differences of condition for all those correct tones at any time point, with the exception of the third tone. For the third tone, there was a significant effect of HNs in which the F0 dip from speech onset to the turning point at Time Point 3 dropped significantly farther than speech alone. This HN dipping effect was not present when all trials were analyzed in Experiment 2a, suggesting that HNs may play a role only when a speaker produces a tone accurately enough for it to be correctly categorized by native judges. Or put another way, for Tone 3 items that were spoken "correctly," HNs served to improve the dipping contour in the tone to make it easier to recognize. Interestingly, hand G's facilitatory effect on the fourth tone in Experiment 2a did not replicate during the F0 analysis of correct-only trials in Experiment 2b. Although this inconstancy between HNs and hand Gs should be interpreted cautiously, one possibility is that these two metaphoric actions affect the speech signal in different ways. We will return to this later.

Given the mostly similar findings for L1 and L2 speakers, we will first discuss the null results from Experiment 2 in the context of foreign language learning, but we will finish by speculating about the two slight but significant effects for the L2 speakers. The L2 null results are consistent with the previous findings that Gs representing phonemic pitch contours (in Mandarin and Japanese) or phonemic length contrasts do not facilitate auditory learning any more than learning with auditory input alone (e.g., Eng et al., 2013; Hirata et al., 2014). In this regard, we apply the same conclusion as that for L1 speakers: Bodily actions, such as hand G, are not naturally designed to play a role at the lowest level, that is, phonemic level, of language processing.

However, the present overall findings contrast with Hannah et al. (2017), Kelly, Bailey, and Hirata (2017), and Morett and Chang (2015), which found robust effects of hand G. Below, we offer four possible explanations for this discrepancy. The first one concerns the nature of the task being different than the present experiment. In Hannah et al. (2017) and Kelly, Bailey, and Hirata (2017), participants had to make *judgments* about what they thought they had perceived. Because these tasks required both perception and judgment, it is possible the results did not reflect the influence of the body on auditory perception but rather the influence of the body on memory and judgment. For Morett and Chang's (2015) findings, the effects of hand Gs were observed in the nonnative speakers' ability to associate Mandarin words with their meaning, whereas Gs had no unique effect on their ability to auditorily identify Mandarin lexical tones.

A second possibility is that because we did not control what participants did when they imitated the model, we may have limited the effects of producing actions on tone production. Indeed, as was mentioned with the L1 speakers, it is possible that the L2 speakers simply produced hand movements that passively followed their speech. In the case of the L2 speakers, this resulted in tones that were not as well formed as L1 speakers, and without video-recording the subjects' own imitated actions, it is not possible to know if the actions themselves were just as sloppy as the speech. That is, perhaps, the L2 speakers did not produce Gs and HNs as dynamically as instructed. Furthermore, unlike previous studies that forced a mismatch between speech and bodily actions (e.g., Hannah et al., 2017; Kelly, Bailey, & Hirata, 2017), our task never forced actions and speech to be in conflict, and that may explain why we did not see an effect of producing actions on speaking tones. Given this limitation, it would be important for future research to control more carefully how L2 learners execute bodily actions while they are producing tones. For example, it would be interesting to force subjects to exaggerate their HNs and hand Gs while articulating the different tones to see if that helps them produce tones that are more in line with L1 speakers. Further, by being able to control and quantify the intensity and range of motor movement, we may be able to see a relationship between movement and performance accuracy.

The third possible explanation may be that behaviors such as HNs and hand Gs are integrated with the phonological system differently in perception than they are in production. There is evidence that the phonological system for speech perception versus speech production only partially overlaps and is not directly linked (e.g., Buchsbaum, Hickok, & Humphries, 2011; Mitterer & Ernestus, 2008). Perhaps, the vocal articulators involved in producing phonemic distinctions—such as tonal contrasts—are relatively encapsulated from bodily actions, whereas the auditory system during perception may be more open to input from other modalities, such as vision (Gentilucci, 2003; McGurk & MacDonald, 1976). This would be interesting because previous research has shown that movements of the hands do affect "higher up" phonological production, such as when a G was made on a specific word over a stretch of utterance (Krahmer & Swerts, 2007). Thus, more research is necessary to better understand the mechanisms that underlie this apparent inconsistency between the roles of the body in speech production versus comprehension.

The fourth possible explanation is that this is not truly a discrepancy: It is possible that our modest effects of HNs and hand Gs in L2 pitch production are consistent with previous research on L2 pitch perception (Hannah et al., 2017; Kelly, Bailey, & Hirata, 2017; Morett & Chang, 2015), but the effects are just much less prominent and widespread in production versus perception. We wish to highlight the fact that, in the case of HNs affecting dipping in the third tone and the case of hand Gs at the descending end of the fourth tone, these effects were observed during the production of low pitches. Moreover, because there were different effects of hand Gs and HNs on different low-dipping tones, it is possible that different metaphoric actions may serve different bodily functions: An exaggerated downward hand motion at the end of a word (as with Tone 4) may facilitate a dropping pitch because of a general release of overall muscle tension (as in Roberge, Kimura, & Kawaguchi, 1996). In contrast, a dipping HN in the middle

of a word (as with Tone 3) might produce particular voice quality changes, such as the well-known "creaky voice" effects seen in Mandarin's third tone (Kuang, 2017). This suggests that different bodily actions may affect different aspects of the speech signal, and future L2 speech training studies should explore these possibilities more systematically.

Differential Effects of Bodily Movements in L1 and L2 Speakers

The L1 and L2 speakers in this study represented two ends of a continuum: On one end, there was complete mastery of the Mandarin tone system, and on the other, there was total inexperience with it. Although the vast majority of our analyses suggested that these two groups were very similar—showing little influence of metaphoric bodily actions on speech production—our L2 group did show two hints of an effect. Hand Gs (Tone 4 in Experiment 2a) and HNs (Tone 3 in Experiment 2b) helped L2 speakers produce tones that finished closer to L1 speakers than the speech baseline. Although these are just two small effects, they are in the predicted direction and deserve some cautious attention. Perhaps, the tonal ability of our L2 speakers was so immature that it was more "open" to influence from other modalities. That is, L2 speakers may have used G and HNs to help bootstrap their vocal system into producing the tones, much in the way that infants move their hands while babbling to give them practice when learning native phonemes (Iverson & Thelen, 1999). In contrast, because our L1 speakers already possessed a mature and fully functioning native phoneme repertoire, their phonological system may have been buffered against other bodily actions produced along with speech.

In this way, it is possible that for L1 speakers, the vocal system leads the way for other motor systems (hands and head), whereas for L2 speakers, nonvocal motor systems can sometimes take more of an active role in directing the vocal system. If this were the case, one might predict that over time, as L2 speakers get more experience with Mandarin, they come to increasingly rely on the other systems to help them master the novel vocalizations, that is, at least until they become true experts, and then, we might expect the system to become more encapsulated such as the current L1 speakers in Experiment 1 (for more on modularization, see Karmiloff-Smith, 1995). Thus, it may be the case for our early L2 speakers in the current study; their perceptual and production system is still immature and, therefore, insufficiently trained to be readily coupled with bodily actions. As a result, we see very few and small effects of Gs or HNs on tonal production in the L2 group. Another prediction is that, if early L2 learners were required to exaggerate their head and hand movements while they spoke different Mandarin tones, these exaggerated movements may drag the vocal system along with it (Roberge et al., 1996). This is not only an interesting possibility for theoretically understanding how different motor systems work together in learning, but it has real practical implications for developing more effective strategies to retune L2 learners' vocal systems to new phonological demands. Because the current study included participants who were at extreme ends of the Mandarin competency spectrum, future studies should focus on students who are in the beginner high, intermediate, or advanced range of fluency. This would help determine whether L2 learners with a more developed, but not quite perfect Mandarin phonological and perceptual system, are able to leverage Gs or HNs to achieve greater production accuracy. Furthermore, having these subjects produce motor movements that are incongruent or congruent with speech with varying degrees of intensity will allow researchers to more fully probe the effect of Gs on tonal production at different stages of language competency.

A "Meta" Function of Metaphoric Bodily Actions

While the results of the acoustic analysis and the native Mandarin-speaking judges' assessments show that hand Gs and HNs did not affect tone production to a significant extent, the questionnaire data suggest that L2 learners believed that a multimodal input—in particular, hand Gs—was quite beneficial. Recall that the majority of L2 speakers in Experiment 2a reported that the G condition helped them to pronounce the tones correctly above any other condition.

What might explain this disconnection? One possibility is that, although hand Gs and HNs do not penetrate down to the phonetic level and influence phonological production to a great degree, it is possible that producing metaphoric bodily actions serves a more metafunction. Perhaps, speaking the tones with Gs and HNs made people feel as if they had done a good job. This possibility makes sense considering the research by Kelly and Goldsmith (2004) on the multiple functions of Gs in learning from a video lecture: They showed that, although viewing Gs did not actually help viewers learn more about the lecture, it did cause them to subjectively report that the lecture was easier to understand (for a related effect of overestimating the helpfulness of pictures in L2 vocabulary learning, see Carpenter & Olson, 2012). Similarly, it may be the case that native English speakers consider videos in which a Mandarin speaker model is demonstrating the tones with bodily actions as getting the point across more clearly, or it could be that the act of imitating the bodily actions in the videos made subjects feel more invested in the task, which, in turn, affected assessments of their own learning. Either way, it suggests an interesting disparity between what bodily actions actually did for subjects and what subjects thought they did.

In general, this issue of what function Gs serve is of growing interest to scholars and teachers. Recently, Kelly, Alibali, and Church (2017) emphasize that Gs can—and almost always do—have overlapping functions. For example, a G can help both a speaker and an addressee during communicative interactions. The present results are consistent with this claim and suggest that Gs can also have multiple functions not only just across speakers and addresses but also within a speaker or addressee. In addition to Gs

(and HNs) having modest benefits for an L2 speaker's tonal production, they may also help this speaker reflect on the experience in a positive way. This has real-life implications. For example, perhaps, mimicking an instructor during a language lesson can make a learner feel more engaged in a language lesson, and that may ultimately function to keep the learner motivated to learn more. This distinction between cognitive and affective functions of bodily actions is important for embodied approaches to language and learning (see Semin & Smith, 2008), and we think the context of foreign language learning is a fertile ground for exploring this issue further.

Conclusion

Considering the results together, the metaphoric bodily actions of hand Gs and HNs seem to have no observable role in the tonal production for L1 speakers and only a modest role for L2 speakers. It is usually dangerous to make too much out of nonsignificant results, but there are good reasons to take them seriously in this case. First, our design had enough power to uncover multiple significant findings with strong effect sizes—including the factors of the tone type and time point—so a lack of power is not the issue. Second, the results converge from different directions: from opposite ends of the Mandarin competence spectrum and by using multiple dependent measures. For these reasons, we are confident in our conclusion regarding the overall minimal role of hand Gs and HNs in Mandarin tone production. Thus, the present results suggest a possible lower limit of the influence of the body on production at the phonemic level for early L2 learners and native L1 speakers.

Amid the excitement of how language and learning are "embodied" processes (Kelly, 2017), it will be increasingly important to identify the boundaries where language remains tethered to the body—and where it breaks free. The present results occupy this gray area. For L2 learning in which the phonological system is not yet locked in such as the adult L1 system, it is possible that metaphoric bodily actions can give a small boost to their learning, which is consistent with Hannah et al. (2017) and Kelly, Bailey, and Hirata (2017). So, although the body may not be optimally designed for lower levels of language, such as processing individual phonemes, it may be co-opted for it if the conditions are right.

Acknowledgments

The authors thank Colgate's Center for Language and Brain for providing funding for participants. The authors would like to thank Youngsun Cho, Eve Yi, Jack Polikoff, Solhee Bae, Yuan Lou, and Ryan Hildebrandt for their efforts in preparing stimuli, running participants and collecting data, and for all their intelligent contributions to and support for the project.

References

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

Boersma, P., & Weenink, D. (2015). Praat: Doing phonetics by computer [Computer program]. Version 5.4.09. Retrieved from http://www.praat.org/

Bolinger, D. (1983). Intonation and gesture. *American Speech, 58*(2), 156–174.

Buchsbaum, B. R., Hickok, G., & Humphries, C. (2001). Role of left posterior superior temporal gyrus in phonological processing for speech perception and production. *Cognitive Science, 25*(5), 663–678.

Burnham, D., Lau, S., Tam, H., & Schoknecht, C. (2001). Visual discrimination of Cantonese tones by tonal but not-Cantonese speakers, and by non-tonal language speakers. In D. Massaro, J. Light, & K. Geraci (Eds.), *Proceedings of the International Conference on Auditory–Visual Speech Processing* (pp. 155–160). Aalborg, Denmark.

Burnham, D., Reynolds, J., Vatikiotis-Bateson, E., Yehia, H., Cioca, V., Haszard Morris, R., . . . Jones, C. (2006). The perception and production of phones and tones: The role of rigid and non-rigid face and head motion. In H. Yehia (Ed.), *Proceedings of the 7th International Seminar on Speech Production* (pp. 1–8). Brazil: CEFALA.

Carpenter, S. K., & Olson, K. M. (2012). Are pictures good for learning new vocabulary in a foreign language? Only if you think they are not. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(1), 92–101.

Casasanto, D., Phillips, W., & Boroditsky, L. (2003). Do we think about music in terms of space? Metaphoric representation of musical pitch. In R. Alterman & D. Kirsch (Eds.), *Proceedings of the 25th Annual Meeting of the Cognitive Science Society* (p. 1323). Austin, TX: Cognitive Science Society.

Chang, Y. S. (2011). Distinction between Mandarin tones 2 and 3 for L1 and L2 listeners. *Proceedings of the 23rd North American Conference on Chinese Lingusitics (NACCL-23), 1,* 84–96.

Chen, C. M. (2013). Gestures as tone markers in multilingual communication. *Research in Chinese as a Second Language, 9,* 143.

Chen, T. H., & Massaro, D. W. (2008). Seeing pitch: Visual information for lexical tones of Mandarin-Chinese. *The Journal of the Acoustical Society of America, 123*(4), 2356–2366.

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.

Eng, K., Hannah, B., Leong, L., & Wang, Y. (2013, June). Can co-speech hand gestures facilitate learning of non-native tones? *Proceedings of Meetings on Acoustics, 19*(1), 060225. https://doi.org/10.1121/1.4799746

Fox, J., & Weisberg, S. (2011). *An {R} companion to applied regression* (2nd ed.). Thousand Oaks, CA: Sage. Retrieved from http://socserv.socsci.mcmaster.ca/jfox/Books/Companion

Gentilucci, M. (2003). Grasp observation influences speech production. *European Journal of Neuroscience, 17*(1), 179–184.

Gentilucci, M., Campione, G. C., Dalla Volta, R., & Bernardis, P. (2009). The observation of manual grasp actions affects the control of speech: A combined behavioral and transcranial magnetic stimulation study. *Neuropsychologia, 47*(14), 3190–3202.

Gluhareva, D., & Prieto, P. (2017). Training with rhythmic beat gestures benefits L2 pronunciation in discourse-demanding situations. *Language Teaching Research, 21*(5), 609–631. https://doi.org/10.1177/1362168816651463

Hannah, B., Wang, Y., Jongman, A., Sereno, J. A., Cao, J., & Nie, Y. (2017). Cross-modal association between auditory and visuospatial information in Mandarin tone perception in noise by native and non-native perceivers. *Frontiers in Psychology, 8,* 2051.

Hirata, Y., & Kelly, S. D. (2010). Effects of lips and hands on auditory learning of second-language speech sounds. *Journal of Speech, Language, and Hearing Research, 53,* 298–310.

Hirata, Y., Kelly, S. D., Huang, J., & Manansala, M. (2014). Effects of hand gestures on auditory learning of second-language vowel length contrasts. *Journal of Speech, Language, and Hearing Research, 57*(6), 2090–2101.

Iverson, J. M., & Thelen, E. (1999). Hand, mouth and brain. The dynamic emergence of speech and gesture. *Journal of Consciousness Studies, 6*(11–12), 19–40.

Jongman, A., Wang, Y., Moore, C., & Sereno, J. A. (2006). Perception and production of Mandarin Chinese tones. In P. Li, E. Bates, L. H. Tan, & O. Tseng (Eds.), *The handbook of East Asian psycholinguistics* (pp. 209–217). Cambridge: Cambridge University Press.

Karmiloff-Smith, A. (1995). *Beyond modularity: A developmental perspective on cognitive science*. Cambridge, MA: MIT Press.

Kelly, S. D. (2017). Exploring the boundaries of gesture–speech integration during language comprehension. In R. B. Church, M. W. Alibali, & S. D. Kelly (Eds.), *Why gesture? How the hands function in speaking, thinking and communicating* (pp. 243–265). Amsterdam, the Netherlands: John Benjamins.

Kelly, S. D., Alibali, M. W., & Church, R. B. (2017). Understanding gesture: Description, mechanism and function. In R. B. Church, M. W. Alibali, & S. D. Kelly (Eds.), *Why gesture? How the hands function in speaking, thinking and communicating* (pp. 3–10). Amsterdam, the Netherlands: John Benjamins.

Kelly, S. D., 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

Kelly, S. D., & Goldsmith, L. (2004). Gesture and right hemisphere involvement in evaluating lecture material. *Gesture, 4,* 25–42.

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,* 673.

Krahmer, E., & Swerts, M. (2007). The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language, 57,* 396–414.

Kuang, J. (2017). Covariation between voice quality and pitch: Revisiting the case of Mandarin creaky voice. *The Journal of the Acoustical Society of America, 142*(3), 1693–1706.

Kuhl, P. K., Conboy, B. T., Padden, D., Nelson, T., & Pruitt, J. (2005). Early speech perception and later language development: Implications for the "critical period." *Language Learning and Development, 1*(3&4), 237–264.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science, 255,* 606–608.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago, IL: University of Chicago Press.

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.

Lenneberg, E. (1967). *Biological foundations of language* (pp. 125–187). New York, NY: Wiley.

Lenth, R. V. (2016). Least-squares means: The R package lsmeans. *Journal of Statistical Software, 69*(1), 1–33.

Lieberman, P., & Blumstein, S. E. (1988). *Speech physiology, speech perception, and acoustic phonetics*. Cambridge, England: Cambridge University Press.

Liu, Y., Wang, M., Perfetti, C. A., Brubaker, B., Wu, S., & MacWhinney, B. (2011). Learning a tonal language by attending to the tone: An in vivo experiment. *Language Learning, 61*(4), 1119–1141.

Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English /r/ and /l/. The role of phonetic environment and talker variability in learning new perceptual categories. *The Journal of the Acoustical Society of America, 94,* 1242–1255.

Loehr, D. (2007). Aspects of rhythm in gesture and speech. *Gesture, 7*(2), 179–214.

McClave, E. (1998). Pitch and manual gestures. *Journal of Psycholinguistic Research, 27,* 69–89.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264,* 746–748.

McNeill, D. (2005). *Gesture and thought*. Chicago, IL: University of Chicago Press.

Mitterer, H., & Ernestus, M. (2008). The link between speech perception and production is phonological and abstract: Evidence from the shadowing task. *Cognition, 109*(1), 168–173.

Morett, L. M., & Chang, L. Y. (2015). Emphasizing sound and meaning: Pitch gestures enhance Mandarin lexical tone acquisition. *Language, Cognition and Neuroscience, 30*(3), 347–353.

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Research report visual prosody and speech intelligibility head movement improves auditory speech perception. *Psychological Science, 15*(2), 133–137.

Oyama, S. (1976). A sensitive period for the acquisition of a non-native phonological system. *Journal of Psycholinguistic Research, 5*(3), 261–283.

Paivio, A. (1986). *Mental representations: A dual coding approach*. Oxford, England: Oxford University Press.

Pisoni, D. B., Aslin, R. N., Perey, A. J., & Hennesy, B. L. (1982). Some effects of laboratory training on identification and discrimination of voicing contrasts in stop consonants. *Journal of Experimental Psychology: Human Perception and Performance, 8,* 297–314.

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Roberge, C., Kimura, M., & Kawaguchi, Y. (1996). *Pronunciation training for Japanese: Theory and practice of the VT method. (in Japanese: Nihongo no Hatsuon Shidoo: VT-hoo no Riron to Jissai)* (pp. 1–185). Tokyo: Bonjinsha.

Semin, G. R., & Smith, E. R. (2008). *Embodied grounding: Social, cognitive, affective, and neuroscientific approaches*. Cambridge: Cambridge University Press.

Tao, L., & Guo, L. (2008). Learning Chinese tones: A developmental account. *Journal of the Chinese Language Teachers Association, 43*(2), 17–46.

Tsao, F. M., Liu, H. M., & Kuhl, P. K. (2004). Speech perception in infancy predicts language development in the second year of life: A longitudinal study. *Child Development, 75*(2), 1067–1084.

Wang, Y., Behne, D., Jongman, A., & Sereno, J. A. (2004). The role of linguistic experiences in the hemisphere processing of lexical tone. *Journal of Applied Psycholinguistics, 25,* 449–466.

Wang, Y., Jongman, A., & Sereno, J. (2003). Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training. *The Journal of the Acoustical Society of America, 113*(2), 1033–1043.

Wang, Y., Jongman, A., & Sereno, J. A. (2006). L2 acquisition and processing of Mandarin Chinese tones. In P. Li, E. Bates, L. H. Tan, & O. Tseng (Eds.), *The handbook of East Asian psycholinguistics* (pp. 250–256). Cambridge: Cambridge University Press.

Wang, Y., Sereno, J., Jongman, A., & Hirsch, J. (2003). fMRI evidence for cortical modification during learning of Mandarin lexical tone. *Journal of Cognitive Neuroscience, 15*(7), 1019–1027.

Wang, Y., Spence, M. M., Jongman, A., & Sereno, J. A. (1999). Training American listeners to perceive Mandarin tones. *The Journal of the Acoustical Society of America, 106,* 3649–3658.

Werker, J., & Curtin, S. (2005). A developmental framework of infant speech processing. *Language Learning and Development, 1,* 197–234.

Werker, J., Fennell, C. T., Corcoran, K. M., & Stager, C. L. (2002). Infant's ability to learn phonetically similar words: Effects of age and vocabulary size. *Infancy, 3,* 1–30.

Werker, J., & Tees, R. (2005). Speech perception as a window for understanding plasticity and commitment in language systems of the brain. *Journal of Developmental Psychobiology, 46,* 233–251.