Multimodality across Epistemologies in Second Language Research

Edited by
Amanda Brown and Søren W. Eskildsen

First published 2024

ISBN: 9781032409818 (hbk)
ISBN: 9781032409832 (pbk)
ISBN: 9781003355670 (ebk)

Chapter 11

Gesture shape and gesture–speech alignment predict simultaneous L2 sound production accuracy

Peng Li, Florence Baills, Xiaotong Xi, and Pilar Prieto

DOI: 10.4324/9781003355670-13

The funder of the Open Access version of this chapter is Spanish Ministry of Science and Innovation.

Routledge
Taylor & Francis Group
NEW YORK AND LONDON

11 Gesture shape and gesture–speech alignment predict simultaneous L2 sound production accuracy

Peng Li, Florence Baills, Xiaotong Xi, and Pilar Prieto

Research focus

Previous studies have suggested that hand gestures can be useful for teaching and learning second language (L2) pronunciation. However, the underlying mechanisms behind this learning process are unclear. This study investigates whether learners' accuracy in imitating gestures during embodied phonetic training predicts their accuracy in producing non-native sounds simultaneously. Twenty-nine Catalan speakers without prior knowledge of Chinese were trained to pronounce Chinese aspirated plosives while doing a fist-to-open-palm gesture mimicking the strong airburst of the aspirated consonants. We assessed the learners' gesture imitation accuracy by gesture shape and the temporal alignment of the gesture and the aspirated sounds. Results showed that more accurate speech imitation (i.e., longer Voice Onset Time (VOT) of the aspirated consonants) was predicted by a more accurate gesture shape and a finer gesture–speech alignment. These results support the importance of the spatiotemporal coupling between hand gestures and speech during embodied L2 phonetic training.

Background

Embodied pronunciation learning

According to the Embodied Cognition paradigm, body and mind are closely related, and more specifically, cognitive processes are grounded in sensory-motor processes in the human body (Ionescu & Vasc, 2014). Body movements can facilitate cognitive activities such as conceptualizing abstract meaning (Barsalou, 2008, 2010) and information recall (Kontra et al., 2015; Mizelle & Wheaton, 2010). The idea that the body plays a direct role in the cognitive process entails important implications for education (see Shapiro & Stolz, 2019 for a summary; see also Tellier, this volume, for the role of the body in language classrooms). In the field of L2 acquisition, it is shown that embodied training involving hand gestures can enhance

DOI: 10.4324/9781003355670-13

the learning of various aspects of L2, including vocabulary (Macedonia, 2014) and grammar (Matsumoto & Dobs, 2017), as well as L2 speech production. For instance, beat gestures cueing speech prominence can improve learners' pronunciation proficiency (Llanes-Coromina et al., 2018), gestures mimicking the nuclear configurations of L2 pitch contours can improve L2 intonation patterns (Yuan et al., 2019), horizontal hand sweep gestures help the production of L2 long vowels (Li et al., 2020), and hand gestures imitating sentence-level prosodic features can contribute to reducing L2 accentedness and improving suprasegmental accuracy (Li et al., 2022). On the learning of specific L2 sounds, gestures, or tactile cues showing acoustic and articulatory facilitated the pronunciation of L2 sounds, such as Spanish /u/ (Hoetjes & van Maastricht, 2020), Chinese aspirated stops (Li et al., 2021; Xi et al., 2020), and English /θ-ð/ (Ozakin et al., 2023) and /æ-ʌ/ contrasts (Xi et al., 2023). Despite these positive findings, the underlying mechanisms that explain the positive effects of embodied training on L2 pronunciation are still not clear.

Moreover, successful embodied phonetic training relies on multiple factors, one of which is the shape of hand gestures. In one of the studies noted above, Li et al. (2021) trained 67 Catalan speakers to learn Chinese aspirated voiceless stops /p^h, t^h, k^h/ by making a fist-to-open-palm hand gesture representing the strong airburst of aspirated stops when producing the sounds. Gesture imitation accuracy was assessed using a holistic rating scheme that integrated both gesture shape and gesture–speech alignment for a given gesture event. Pronunciation accuracy was assessed in a pretest/posttest/delayed posttest paradigm by measuring the learners' VOT of the plosives. The results revealed that while good gesture performers showed an improvement after training and had maintained it after three days (the delayed post-test), poor gesture performers showed no significant change in VOT across the three tests.

Furthermore, if the learners observe hand gestures that do not effectively represent the target phonetic features, gestural training will lead to null or even negative effects on L2 pronunciation. For example, Hoetjes and van Maastricht (2020) found that using a complex iconic gesture shape could even harm the learning of non-native interdental fricatives, possibly because "seeing the iconic gesture cost a fair amount of processing energy" (p. 14). Similarly, although Xi et al. (2020) found that having learners observe the same gesture as Li et al. (2021) helped the learning of aspirated plosives, it had no such effect on aspirated affricates, probably because the gesture shape did not accurately mimic the prolonged friction period of aspirated affricates.

Another factor that may potentially affect the effects of L2 phonetic training is the temporal alignment of speech and gesture. Crosslinguistic research has shown evidence that gesture and human speech are temporally

aligned (see Shattuck-Hufnagel, 2019 for a review). For example, pointing gesture apexes – the phase of the gesture that contains the kinematic peak of maximum velocity – frequently co-occur with peaks in intonation (Esteve-Gibert & Prieto, 2013) and speech prominence (Danner et al., 2018). Beat strokes are aligned with pitch-accented syllables (Shattuck-Hufnagel & Ren, 2018). The maximum extension of arm movements co-occurs with peaks in pitch and amplitude (Pouw et al., 2020). Interestingly, the arm movement affects airflow during respiration, which in turn influences vocalization, and modifies the pronunciation of individual sounds and syllables (Pouw et al., 2020; Pouw et al., 2021). Nevertheless, despite the cross-linguistic evidence enumerated above, little research in L2 speech training has focused on the role of gesture–speech alignment in L2 phonetic training.

Goals of the study

Based on the literature review, we surmise that accuracy in the imitation of the gesture and the alignment with speech during L2 embodied phonetic training may condition the role of hand gestures in embodied phonetic training. Therefore, this chapter further investigates the importance of accurate gesture imitation during phonetic training. The database for this chapter comes from Li et al.'s (2021) training experiment. Li et al.'s study assessed the effects of the participants' gesture imitation accuracy during an embodied training session which focused on the improvement after training. By contrast, in this chapter, we analyzed the participants' speech and gesture performance *during* this training session. Therefore, the novelty of this research is to test the claim that gesture imitation accuracy does not only affect pronunciation gains *after* training, which was proven by Li et al.'s study, but also the simultaneous speech production accuracy *during* training.

We address the following research question in this chapter: To what extent do learners' accuracy in gesture shape and the temporal alignment between the gesture and target L2 sounds predict the pronunciation accuracy of the target sounds? To answer this question, we analyzed the performance of 29 learners during a five-minute embodied training experiment reported by Li et al. (2021). The target L2 sounds were the Chinese aspirated stops /p^h, t^h, k^h/, which are not part of the sound inventory of the participants' native language, Catalan.

The participants' gesture imitation accuracy during training was assessed based on two complementary aspects, namely (a) the accuracy of the gesture shape and (b) the precision of gesture–speech temporal alignment. In terms of gesture shape, as the gesture was a fist-to-open-palm movement, we decomposed the gesture into two phases, the "fist" phase, and the "palm" phase. Then we evaluated whether the participants imitated each phase

accurately and measured the duration of each phase. As for gesture–speech alignment, we analyzed the temporal distance in milliseconds between the onset of aspiration and the moment when the palm began to open.

We hypothesized that both accurate gesture shape and gesture–speech alignment would significantly correlate with the learners' simultaneous speech imitation accuracy. Specifically, participants who held their fists closed and then opened their hands precisely at the moment of aspiration would better channel their energy into producing a strong airburst, resulting in more native-like aspirated stops. As for gesture–speech alignment, we predicted that by opening their palms *no later than* the airburst release, participants would produce an airburst with sufficient strength, resulting in a more accurate pronunciation of the target aspirated sounds.

Method

Participants

Twenty-nine Catalan speakers (26 females; age 18–24 years, M = 19.31, SD = 1.70), who had no prior knowledge of Chinese, voluntarily participated in the experiment. All the participants gave written consent, allowing the researchers to record and analyze the oral and gestural data collected during the experiment.

Materials

The training materials were six pairs of Chinese disyllabic words contrasting only in the word-initial aspiration of the plosive consonant (e.g., *tu li* 'independence' vs *t^{h}u li* 'legend', see Li et al., 2021 for full details). Participants watched videos of two native Chinese-speaking instructors (one female) producing the 12 target words. The two instructors mimicked the airburst using their hands when producing the aspirated plosives. They held their fists firmly, raised both fists to the height of their shoulders, and quickly opened their palms toward the camera during the aspirated consonants to visually represent the airburst (refer to Li et al., 2021). No gesture was performed for unaspirated /p, t, k/; therefore, these sounds were not part of the analysis in this chapter.

Procedure

Before undergoing training, participants received a brief introduction regarding the Chinese aspirated consonant contrasts and instructions about what they would be expected to do during the session. They were explicitly instructed to repeat the training words after the instructors and simultaneously imitate the gestures they had seen.

Participants did the experiment individually in a silent room and were video recorded throughout the session using the laptop camera, which was also used to present the stimuli. During the training session, each of the training words was trained four times, with the unaspirated word always appearing before the aspirated one. In each of the four trials, participants first saw the approximate phonetic transcription of the target word adapted to the Catalan orthography, together with a superscripted "h" to indicate the aspirated consonants (e.g., *t^{h}u li*). Next, they saw an instructor uttering the two target words consecutively. They also saw the instructor perform the target hand gestures for the words containing the aspirated consonants. After hearing the instructor, a three-second black screen with the instruction *Repeteix-ho* 'Repeat that' appeared, to remind participants to repeat both the speech and gestures. The training session lasted around five minutes altogether. All verbal and written instructions were given in Catalan to ensure full understanding.

Data coding

The data for analysis in this chapter were extracted from the video recordings of the 29 participants. The total duration of the audiovisual corpus was 145 min (29 participants × 5 minutes per training session). From the video files, we extracted a total of 696 clips showing participants producing target words while gesturing (29 participants × 6 aspirated items × 4 repetitions). Out of the total clips extracted, seven were excluded as the participants' gesturing position was too low for the camera to capture all gesture movements. The remaining 689 clips were then annotated as follows.

a. Speech annotation. The first author annotated the VOT of the aspirated consonants produced by the participants using Praat (Boersma & Weenink, 2020). VOT measures assess the time interval between the airburst and the onset of voicing (Johnson, 2011). Chinese aspirated plosives /p^{h}, t^{h}, k^{h}/ show a mean VOT of around 114.40 ms in isolated words (Xi et al., under review), which is much longer than the VOT reported for the Catalan voiceless plosive counterparts /p, t, k/, which ranges from 0 to 30 ms (Aliaga-García & Mora, 2008). Therefore, since the gestures were designed to mimic the strong airburst, performing them was expected to trigger longer VOT in the participants' online speech imitation than what they would naturally produce in their native language. The audio data were then imported from the Praat Text Grid into ELAN (*ELAN*, 2022), where the aspiration phase of the target sounds was annotated as "VOT".

b. Gesture annotation. The first author manually annotated two sets of information on the Gesture tier, namely *Gesture shape* and *Gesture–speech alignment*.

(1) Gesture shape. We defined two phases for each gesture event, the *fist* and *palm* phases. We annotated each gesture in a binary coding system (two levels, "accurate" or "inaccurate") by visually assessing the accuracy with which learners produced the fist and the palm phases, respectively. For the fist phase, if the learner's fingers were all bent and bunched tightly together on the palms, it was labeled as an "accurate" fist. All other shapes, such as not firmly bunching the fingers on the palms or not bending the fingers at all, were labeled as "inaccurate". Likewise, a palm gesture was labeled as "accurate" if, by the end of the opening gesture, all five fingers were fully extended outwards. By contrast, an "inaccurate" palm denoted all other behaviors, such as slowly spreading the fingers or incompletely opening the hands. Figure 11.1 shows examples of accurate and inaccurate fist and palm gestures.

(2) Gesture–speech alignment. First, we annotated the starting and ending points of the fist and palm phases on the Gesture tier. Then, a total of four measures were exported from ELAN, namely (a) the duration of the fist phase (henceforth "fist duration"), (b) the fist shape (two levels: Accurate or inaccurate), (c) the duration of the palm phase (henceforth "palm duration"), and (d) the palm shape (two levels: Accurate or inaccurate). Figure 11.2 shows a screenshot of an ELAN page with annotations of gesture and speech events.

Finally, we also exported the starting point of the aspiration from the Speech tier and the onset of the palm-opening gesture from the Gesture tier to measure the time elapsed between the two points. This variable, referred to as "gesture latency", serves as an index of gesture–speech alignment (Gesture latency = Time $_{\text{VOT onset}}$ – Time $_{\text{palm opening onset}}$). Hence, a *positive* gesture latency means that the sound started *after* opening the palm, while a negative value indicates the opposite. See Figure 11.2 for a gesture event with positive gesture latency. *Statistical analysis*

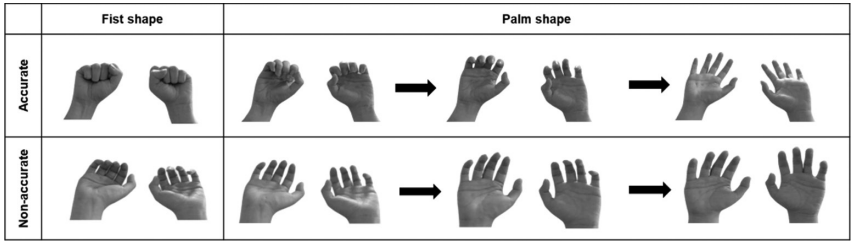

Figure 11.1 Upper left: Accurate fist shape produced with both hands tightly clenched. Upper right: Accurate palm shape with five fingers of both hands quickly extended outward. Lower left: Inaccurate fist shape with the fingers not closed into fists. Lower right: Inaccurate palm shape with fingers not fully extended.

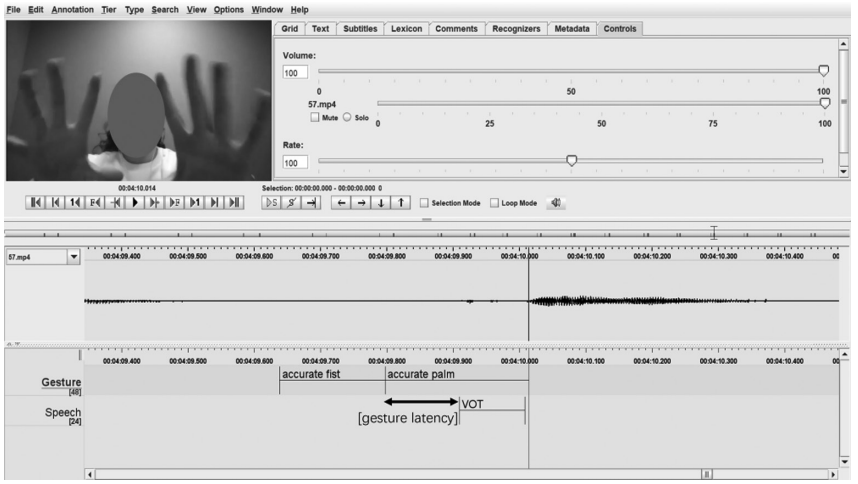

Figure 11.2 Screenshot of an ELAN annotation page. The Gesture tier shows the duration and accuracy rate of each gesture phase, the fist and palm phases. The Speech tier shows the VOT of aspirated consonants imported from the Praat Text Grid. The tag "gesture latency" is added to the screenshot to show the time lag between the initiation of the palm phase of the gesture and the VOT.

We built a linear mixed effects model (LMM) using the *lme4* package (Bates et al., 2015) in R (R Core Team, 2014) to model the data. First, all continuous variables, including VOT, fist duration, palm duration, and gesture latency, were normalized using *z*-scores. The dependent variable of the model was the VOT. The fixed effects included all five measures exported from ELAN: Fist duration, fist shape, palm duration, palm shape, and gesture latency. We also included two two-way interactions: Fist Duration × Fist Shape and Palm Duration × Palm Shape. Second, we fitted the model with the most complex random structure, and the best-fitting random structure was determined using the *buildmer()* function from the *buildmer* package (Voeten, 2021), which identified random intercepts for participants and items, by-participant random slopes for gesture latency and palm shape, and by-item random slopes for fist shape and palm shape.

Finally, we calculated the significance using the Type II Wald *chi*-squared test from the *car* package (Fox & Weisberg, 2019) and carried out post-hoc comparisons using the *emtrends()* function with the significance values adjusted by the Bonferroni method using the *emmeans* package (Lenth et al., 2020).

Findings

The descriptive results are summarized in Table 11.1 from the by-item data set which was used for statistical analyses.

Additionally, we calculated the by-participant accuracy rate of the fist shape and palm shape. Overall, the mean accuracy rate of the fist phase was 69.81% (*SD* = 36.36) and 74.80% (*SD* = 34.50) for the palm phase.

Effects of gesture shape and duration

The LMM analysis revealed a significant main effect of fist duration, $\chi^2(1) = 17.33$, $p < 0.001$, with longer fist duration relating to longer VOT, $\beta = 0.24$, $SE = 0.05$, $t = 5.35$. However, although there was a significant main effect of fist shape, $\chi^2(1) = 6.73$, $p = 0.009$, the post-hoc analysis did not show significant differences in the VOT produced with different fist shape (i.e., accurate vs inaccurate) after Bonferroni adjust for the *p* values, $\Delta\beta = 0.30$, $t(7.9) = 1.85$, $SE = 0.16$, $p = 0.102$.

More importantly, there was a significant two-way interaction of Fist Duration × Fist Shape, $\chi^2(1) = 11.35$, $p = .001$, which suggests that the effects of fist duration on VOT varied as a function of fist shape. Post-hoc comparisons revealed that fist duration significantly affected VOT only when the fist shape was accurate, $\beta = 0.24$, $SE = 0.05$, with longer fist duration triggering longer VOTs. For the inaccurate fist shape, fist duration could not significantly affect the VOT of the target consonant, $\beta = 0.03$, $SE = 0.05$. The difference in the coefficients between accurate and inaccurate fist shape was significant, $\Delta\beta = 0.22$, $SE = 0.07$, $t(550) = 3.23$, $p = 0.001$. By contrast, no significant main effects were found for palm duration, $\chi^2(1) = 0.73$, $p = 0.392$, palm shape, $\chi^2(1) = 0.08$, $p = 0.773$, or the two-way interaction of Palm Duration × Palm Shape, $\chi^2(1) = 0.10$, $p = 0.753$. These results suggest that palm duration and palm shape did not significantly affect the VOT value of the co-occurring target sounds.

Table 11.1 Count (N), mean (M), standard deviation (SD), and range of the duration of fist shape, palm shape, gesture latency, and VOT in milliseconds.

	N	*M*	*SD*	*Range*
Fist shape	689	144.85	56.51	48–478
Accurate fist	482	137.93	49.51	51–332
Inaccurate fist	207	160.97	67.58	48–478
Palm shape	689	195.39	46.7	68–440
Accurate palm	518	188.37	40.09	68–440
Inaccurate palm	171	216.65	57.74	114–412
Gesture latency	689	160.96	145.77	–427–945
VOT	689	91.73	39	6–232

Effect of gesture–speech alignment

There was a significant main effect of gesture latency, $\chi^2(1) = 30.37$, $p < 0.001$, which indicates that the temporal alignment between gesture and speech could significantly predict the VOT duration, $\beta = 0.45$, $SE = 0.09$. The positive coefficient suggests that when the gesture latency increased, the VOT produced by participants also increased. Thus, the participants produced longer VOTs when they released the airburst *after* they began to open their palms rather than *before* they did so, and a longer gesture latency after the palm opening was more likely to trigger a longer VOT.

Figure 11.3 visually plots the two-way interaction between Fist Duration × Fist Shape (left panel) and the significant main effect of gesture latency (right panel). Note that all the values shown in Figure 11.3 are plotted in their original units (milliseconds), not their *z*-scores, to facilitate the readability of the results. In the right panel, the vertical dashed line indicates the point at which the aspirated consonant is produced at exactly the moment when the palm is about to open (gesture latency = 0), and a positive value indicates that the sound starts being produced after the palm has started to open. Presumably, because the firmly clenched fist accumulated considerable energy, the tension in the hands may have affected the strength of the airburst. Therefore, producing the target sounds after palm opening (gesture latency > 0) indicated a more accurate gesture–speech alignment pattern.

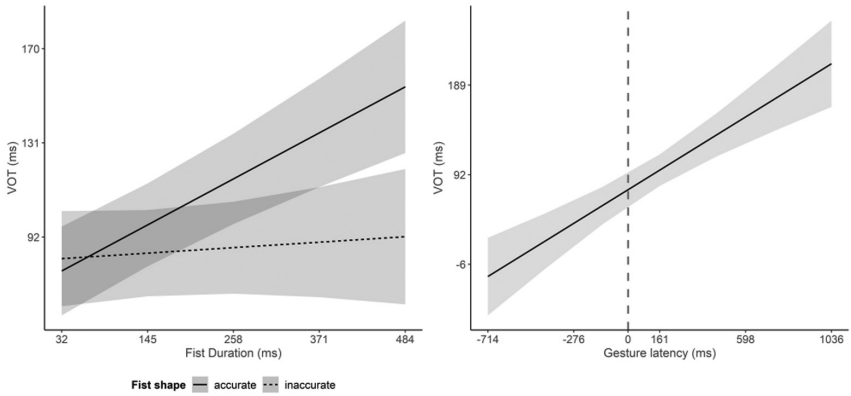

Figure 11.3 LMM analysis of the interaction between fist duration and fist shape (left panel) and gesture latency (right panel), both plotted against estimated VOT values. In both panels, the gray bands indicate the 95% confidence intervals.

Conclusion

This chapter investigated how gesture imitation accuracy during an embodied L2 phonetic training session affected the accuracy of simultaneous speech production. Twenty-nine Catalan speakers were trained to pronounce non-native Chinese aspirated plosives by repeating Chinese words after an instructor while simultaneously imitating the instructor's fist-to-open-palm gesture, a hand gesture intended to represent the strong airburst of the target aspirated sounds. Participants' gesture imitation accuracy was assessed in terms of hand gesture shape and temporal alignment between the palm-opening and the start of aspiration. Participants' pronunciation accuracy of the target aspirated plosives was evaluated by measuring the VOT, with longer VOT indicating more target-like pronunciation.

In terms of gesture shape, the results showed that the shape of the fist phase of the gesture movement (the fist shape) and its duration were important to predict the simultaneous L2 sound production accuracy. When the learners could accurately hold their fist with both hands tightly clenched, a longer fist duration triggered longer VOT, indicating the more accurate pronunciation of the target aspirated sounds. This shows that holding the fists firmly was an important embodiment cue for learners. This body metaphor and holding the fist for a long time may have induced the accumulation of more energy to generate a stronger airburst, resulting in more accurate VOT. By contrast, the duration of an inaccurate fist shape (e.g., not holding the fists firmly, or not holding the fists at all) did not show a significant effect on the VOT values, indicating that an erroneous fist shape, even if held longer, did not necessarily entail the production of a longer VOT. Finally, no significant effects of either palm shape or palm duration were found on VOT, suggesting that neither of these gesture features was relevant for the pronunciation of aspirated sounds.

Regarding gesture–speech alignment, the results showed that a tight gesture–speech alignment, as reflected by the gesture latency between the palm opening phase and the start of the airburst of the aspiration, positively affected simultaneous speech production accuracy. Specifically, an airburst released right *after* the palm opening was stronger than an airburst released *before* the palm opening. It is easy to conjecture that a sudden opening of the firmly held fists would prepare the speaker to release a stronger airburst before the articulation, and it thus can be interpreted as a tighter gesture–speech alignment pattern. Since hand movements temporally align with speech acoustics via the respiratory-vocal system (Pouw, et al., 2020), a palm-opening gesture produced after the release would point to inaccurate alignment because the accumulated energy has already been released before the motor movement.

More importantly, our results complement and expand previous results on the positive correlations between learners' gesture imitation accuracy and their learning outcomes as reported by Li et al. (2021). The results reported here suggest that during the training phase, learners' on-target

VOT had already varied depending on how accurately they performed the gestures, which explains why gesture performance accuracy was important for training the pronunciation of the Chinese L2 sounds. Interestingly, participants showed considerable individual differences in terms of gesture shape accuracy as reflected by the large standard deviation (see Findings section). In our view, individual differences in flexible motor-sensory control and attention control may have affected the learners' gesture performance during training. Future studies could explore which individual factors would affect the shape of gestures.

In conclusion, this study highlights the importance of learners' abilities to imitate gestures accurately during embodied L2 phonetic training. First, the results offer supporting evidence in L2 for recent theories underscoring the spatiotemporal coupling between hand gestures and speech, more specifically between manual and articulatory movements (Gentilucci & Dalla Volta, 2008; Parrell et al., 2014). Second, it provides direct evidence for the embodied cognition paradigm (Barsalou, 2008, 2010) by showing that gesture and speech constitute an integrated system in embodied L2 phonetic training. All in all, the tight relationship between speech and gesture offers a unique opportunity for the development and testing of embodied methodologies in L2 pronunciation teaching and learning.

Acknowledgments

This study is funded by "Multimodal Communication: The integration of prosody and gesture in human communication and in language learning" [PID2021-123823NB-I00] awarded by the Ministerio de Ciencia e Innovación and "Multimodal language learning: Prosodic and Gestural Integration in Pragmatic and Phonological Development" [PGC2018-097007-B-I00], awarded by the Ministerio de Ciencia, Innovación y Universidades, Agencia Estatal de Investigación, and Fondo Europeo de Desarrollo Regional. PL is supported by the Research Council of Norway through its Centres of Excellence funding scheme [223265]. FB acknowledges a Margarita Salas grant funded by European Union–NextGenerationEU, Ministry of Universities and Recovery, Transformation, and Resilience Plan, through a call from Pompeu Fabra University. XX is supported by the Secretaria d'Universitats i Recerca de la Generalitat de Catalunya and the European Social Fund under the Grant for the recruitment of early-stage research staff [2021FI_B 00137].

Open Access publication of this chapter was funded by 'Multimodal Communication: The integration of prosody and gesture in human communication and in language learning' (PID2021-123823NB-I00) awarded by Spanish Ministry of Science and Innovation.

References

Aliaga-García, C., & Mora, J. C. (2008). Perception and production of oral stops by Catalan/Spanish learners of English: A phonetic training experiment. In R. Monroy Casas & A. Sánchez Pérez (Eds.), *25 Años de Lingüística en España. Hitos y retos* (pp. 9–15). Universidad de Murcia. http://www.um.es/lacell/aesla/contenido/pdf/1/aliaga.pdf

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, *59*, 617–645. https://doi.org/10.1146/annurev.psych.59.103006.093639

Barsalou, L. W. (2010). Grounded cognition: Past, present, and future. *Topics in Cognitive Science*, *2*(4), 716–724. https://doi.org/10.1111/j.1756-8765.2010.01115.x

Bates, D., Mächler, 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. (2020). *Praat:* Doi*ng phonetics by computer* [Computer Program]. Retrieved January 6, 2016, from http://www.praat.org

Danner, S. G., Barbosa, A. V., & Goldstein, L. (2018). Quantitative analysis of multimodal speech data. *Journal of Phonetics*, *71*, 268–283. https://doi.org/10.1016/j.wocn.2018.09.007

ELAN (Version 6.4). (2022). *Max Planck institute for psycholinguistics, the language archive*. https://archive.mpi.nl/tla/elan

Esteve-Gibert, N., & Prieto, P. (2013). Prosodic structure shapes the temporal realization of intonation and manual gesture movements. *Journal of Speech, Language, and Hearing Research*, *56*(3), 850–864. https://doi.org/10.1044/1092-4388(2012/12-0049)

Fox, J., & Weisberg, S. (2019). *An {R} companion to applied regression* (3rd ed.). Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Gentilucci, M., & Dalla Volta, R. (2008). Spoken language and arm gestures are controlled by the same motor control system. *Quarterly Journal of Experimental Psychology*, *61*(6), 944–957. https://doi.org/10.1080/17470210701625683

Hoetjes, M., & van Maastricht, L. (2020). Using gesture to facilitate L2 phoneme acquisition: The importance of gesture and phoneme complexity. *Frontiers in Psychology*, *11*(03178), 1–16. https://doi.org/10.3389/fpsyg.2020.575032

Ionescu, T., & Vasc, D. (2014). Embodied cognition: Challenges for psychology and education. *Procedia - Social and Behavioral Sciences*, *128*, 275–280. https://doi.org/10.1016/j.sbspro.2014.03.156

Johnson, K. (2011). *Acoustic and auditory phonetics* (3rd ed.). Blackwell Publishing.

Kontra, C., Lyons, D. J., Fischer, S. M., & Beilock, S. L. (2015). Physical experience enhances science learning. *Psychological Science*, *26*(6), 737–749. https://doi.org/10.1177/0956797615569355

Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2020). *Emmeans: Estimated marginal means, Aka Least-Squares means*. R package. https://cran.r-project.org/package=emmeans

Li, P., Baills, F., Baqué, L., & Prieto, P. (2022). The effectiveness of embodied prosodic training in L2 accentedness and vowel accuracy. *Second Language Research*. https://doi.org/10.1177/02676583221124075

Li, P., Baills, F., & Prieto, P. (2020). Observing and producing durational hand gestures facilitates the pronunciation of novel vowel-length contrasts. *Studies in Second Language Acquisition*, *42*(5), 1015–1039. https://doi.org/10.1017/S0272263120000054

Li, P., Xi, X., Baills, F., & Prieto, P. (2021). Training non-native aspirated plosives with hand gestures: Learners' gesture performance matters. *Language, Cognition and Neuroscience*, *36*(10), 1313–1328. https://doi.org/10.1080/23273798.2021.1937663

Llanes-Coromina, J., Prieto, P., & Rohrer, P. (2018). Brief training with rhythmic beat gestures helps L2 pronunciation in a reading aloud task. *Proceedings of the speech prosody*, *2018*, 498–502. https://doi.org/10.21437/SpeechProsody.2018-101

Macedonia, M. (2014). Bringing back the body into the mind: Gestures enhance word learning in foreign language. *Frontiers in Psychology. DEC*, *5*, 1–6. https://doi.org/10.3389/fpsyg.2014.01467

Matsumoto, Y., & Dobs, A. M. (2017). Pedagogical gestures as interactional resources for teaching and learning tense and aspect in the ESL grammar classroom. *Language Learning*, *67*(1), 7–42. https://doi.org/10.1111/lang.12181

Mizelle, J. C., & Wheaton, L. A. (2010, December). Why is that hammer in my coffee? A multimodal imaging investigation of contextually based tool understanding. *Frontiers in Human Neuroscience*, *4*, 1–14. https://doi.org/10.3389/fnhum.2010.00233

Ozakin, A., Xi, X., Li, P., & Prieto, P. (2023). Thanks or tanks: Training with tactile cues improves learners' accuracy of English interdental consonants in an oral reading task. *Language Learning and Development*, *19*(4), 404–419. https://doi.org/10.1080/15475441.2022.2107522

Parrell, B., Goldstein, L., Lee, S., & Byrd, D. (2014). Spatiotemporal coupling between speech and manual motor actions. *Journal of Phonetics*, *42*, 1–11. https://doi.org/10.1038/jid.2014.371

Pouw, W., de Jonge-Hoekstra, L., Harrison, S. J., Paxton, A., & Dixon, J. A. (2021). Gesture–speech physics in fluent speech and rhythmic upper limb movements. *Annals of the New York Academy of Sciences*, *1491*(1), 89–105. https://doi.org/10.1111/nyas.14532

Pouw, W., Harrison, S. J., & Dixon, J. A. (2020). Gesture-speech physics: The biomechanical basis for the emergence of gesture-speech synchrony. *Journal of Experimental Psychology: General*, *149*(2), 391–404. https://doi.org/10.1037/xge0000646

Pouw, W., Harrison, S. J., Esteve-Gibert, N., & Dixon, J. A. (2020). Energy flows in gesture-speech physics: The respiratory-vocal system and its coupling with hand gestures. *The Journal of the Acoustical Society of America*, *148*(3), 1231–1247. https://doi.org/10.1121/10.0001730

R Core Team. (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.r-project.org/

Shapiro, L., & Stolz, S. A. (2019). Embodied cognition and its significance for education. *Theory and Research in Education*, *17*(1), 19–39. https://doi.org/10.1177/1477878518822149

Shattuck-Hufnagel, S. (2019). Toward an (even) more comprehensive model of speech production planning. *Language, Cognition and Neuroscience*, *34*(9), 1202–1213. https://doi.org/10.1080/23273798.2019.1650944

Shattuck-Hufnagel, S., & Ren, A. (2018, September). The prosodic characteristics of non-referential co-speech gestures in a sample of academic-lecture-style speech. *Frontiers in Psychology*, *9*, 1–13. https://doi.org/10.3389/fpsyg.2018.01514

Voeten, C. C. (2021). *Buildmer: Stepwise elimination and term reordering for mixed-effects regression*. https://cran.r-project.org/package=buildmer

Xi, X., Li, P., Baills, F., & Prieto, P. (2020). Hand gestures facilitate the acquisition of novel phonemic contrasts when they appropriately mimic target phonetic features. *Journal of Speech, Language, and Hearing Research*, *63*(11), 3571–3585. https://doi.org/10.1044/2020_JSLHR-20-00084

Xi, X., Li, P., Baills, F., & Prieto, P. (Under review). Appropriate gesture performance helps the pronunciation of non-native segmental features more than gesture observation. *Studies in Second Language Learning and Teaching*.

Xi, X., Li, P., & Prieto, P. (2023). Reducing acoustic overlap of L2 English vowels through gestures encoding lip aperture. In A. Henderson & A. Kirkova-Naskova (Eds.), *Proceedings of the 7th conference of English pronunciation: Issues & Practices: Issues and Practices* (pp. 261–270). Université Grenoble-Alpes. https://doi.org/10.5281/zenodo.8225191

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