

Review

A Review of the Effectiveness of Hand Gestures in Second Language Phonetic Training

Xiaotong Xi ¹ and Peng Li ^{2,*}

¹ School of Foreign Studies, Shandong University of Finance and Economics, Jinan 250014, China; xiaotong.xi@sdufe.edu.cn

² Center of Linguistics, School of Arts and Humanities, University of Lisbon, 1600-214 Lisbon, Portugal

* Correspondence: peng.li@edu.ulisboa.pt

Abstract

This narrative review synthesizes 24 empirical studies on the role of four types of pedagogical gestures (beat, durational, pitch, and articulatory) in second language (L2) phonetic training since 2010. We reviewed studies involving training interventions to assess the efficacy, mediating factors, and robustness of multimodal training. The findings confirm that gestural training is a powerful tool, yielding the most robust positive effects for L2 speech production and the acquisition of suprasegmental features. Crucially, the effectiveness is highly dependent on gesture-sound consistency and visual saliency of the target phonetic/prosodic feature. However, results are mixed regarding perceptual learning and the generalization of gains to untrained items or novel contexts. While the literature supports the value of gestural training, there are gaps in determining the optimal training paradigm (observing gestures vs. performing gestures), accounting for individual learner differences, and establishing long-term retention and ecological validity. Future research should incorporate longitudinal designs and neurophysiological methods to fully illuminate the cognitive mechanisms that drive the body–mind link in L2 speech acquisition.

Keywords: embodied cognition; hand gestures; second language acquisition; speech perception; speech production; multimodal phonetic training

1. Introduction

In second language (L2) speech acquisition, adult learners often struggle with non-native sounds and prosodic features. A wide range of techniques has been explored to improve L2 speech learning, among which gestural training has gained increasing attention (Gullberg, 2022). As a subtype of multimodal training, it involves the use of pedagogical gestures, which encompass both co-speech gestures and speech-linked gestures that teachers often use in classrooms to support the learning of both segmental and suprasegmental features. These gestures are particularly accessible because they do not require specialized technology and can be easily integrated into classroom instruction (Li et al., 2023a).

The effectiveness of gestural training is grounded in the framework of Embodied or Grounded Cognition (Matheson & Barsalou, 2018). This perspective argues that body and cognition share the same processing system (Ionescu & Vasc, 2014), and, as a result, bodily actions can influence cognitive processing (Wilson, 2002). The close relationship between body and mind has clear educational implications. As noted by Shapiro and Stolz (2019), embodiment can support learning by enriching cognitive processing or acting as a diagnostic tool to evaluate conceptual understanding. In practice, hand gestures can benefit



Academic Editor: Donna Erickson

Received: 15 December 2025

Revised: 19 February 2026

Accepted: 25 February 2026

Published: 4 March 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

both teachers and learners. Teachers can infer students' comprehension by observing their gestures, while learners' own gestures can transfer some cognitive load from verbal to visuospatial resources, which facilitates learning (Sweller et al., 1998). Empirical studies have similarly shown that bodily actions can aid abstract conceptualization (Barsalou, 2008, 2010), enhance recall (Kontra et al., 2015; Mizelle & Wheaton, 2010), and serve as a natural method for cognitive offloading (Wilson, 2002).

According to McNeill (1992), hand gestures can be categorized into four types: (a) Iconic gestures refer to spatial hand movements that depict objects or actions. (b) Metaphorical gestures involve hand movements that visualize abstract concepts. (c) Deictic gestures (also called "pointing gestures") are hand movements that point to a specific object or direction. (d) Beat gestures are hand movements that do not encode semantic meaning but often accompany prominence in speech. All four types of gestures can be used in L2 learning. As iconic and metaphorical gestures can represent semantic meaning, early research has shown that they can help with word learning (Allen, 1995; Kelly et al., 2009; Kelly & Lee, 2012; Macedonia, 2014; Tellier, 2008). Later studies extended this line of research to L2 speech learning, focusing on phonetic aspects. Because iconic and metaphorical hand gestures can illustrate a specific object, action, or abstract concept, they can be used to represent the acoustic and prosodic properties of L2 speech. Pointing gestures can guide learners' attention to specific articulatory movements to help them process learning targets. Beat gestures, by their alignment with prosodic structure, can highlight speech prominence, which is useful for L2 phonetic training. Although not tied to lexical meaning, these gestures are useful pedagogical tools in L2 phonetic training. Therefore, in this review article, we use the broad term "pedagogical gestures" to refer to hand gestures designed to illustrate L2 phonetic and prosodic features for training purposes. Because gesture and speech are closely linked in first language (L1) speech (Iverson & Goldin-Meadow, 2005; McNeill, 1992; Yang & Shu, 2016), understanding how gestures shape L2 speech learning is an important goal for multimodal research. Since the early 2010s, a considerable number of studies have examined the role of gestures in L2 phonetic training. Therefore, this review focuses on studies published after 2010. These studies have empirically assessed pedagogical gestures targeting specific phonetic features, either by highlighting prosodic prominence or depicting other target phonetic features such as duration, pitch, and articulatory features. Accordingly, the diverse types of pedagogical gestures can be divided into four major types, each targeting a specific phonetic or prosodic property: beat gestures, durational gestures, pitch gestures, and gestures illustrating articulatory features.

Although evidence is mounting that pedagogical gestures facilitate L2 speech learning, the results are mixed. To solve the puzzle of the mixed results, the current review focuses on three aspects. First, there are two paradigms in gestural training: instructors can either have learners actively perform pedagogical gestures or merely observe them. If embodiment can support learning, performing hand gestures should outperform observing gestures without enacting them. However, empirical studies have not found consistent evidence to support this hypothesis (e.g., Baills et al., 2019; Hirata et al., 2014). Therefore, synthesizing training studies with hand gestures will help identify the potential factors that influence the effectiveness of such training. Second, the outcomes of phonetic training can be measured in two modalities: perception and production. Recent research suggests that L2 speech perception and production may co-evolve in a non-parallel fashion (Flege & Bohn, 2021; Nagle & Baese-Berk, 2022). In this sense, gestural training offers a way to observe this perception-production relationship from a developmental perspective. However, empirical evidence regarding the effects of gestural training remains inconsistent across these modalities (e.g., Li et al., 2020; Xi et al., 2020), further highlighting the complex interplay between perception and production in L2 speech acquisition. Therefore, the second goal of this review is to synthesize how and why pedagogical gestures facilitate perceptual and productive learning in different ways. Third, a

robust training technique should be effective not only in terms of immediate gains but also in terms of generalization and retention effects (Logan & Pruitt, 1995; Rato & Oliveira, 2023). Immediate gains are often measured by comparing learners' performance before the training (pretest) and immediately after the training (posttest). Generalization effects can be measured by assessing whether learners can generalize their knowledge to untrained items or new contexts. Retention effects are often measured via a delayed posttest implemented after the training in a given period of time. We also focus on synthesizing these effects of pedagogical gestures from the existing literature. With these three main purposes, this review summarizes the findings of training studies involving four major types of pedagogical gestures that have been most widely investigated in the literature: beat gestures, durational gestures, pitch gestures, and gestures illustrating articulatory features. We then discuss the synthesized results, identify gaps in the current research, and propose directions for future work.

As this review focuses on training studies, two inclusion criteria have been established. First, the study should provide at least one training method with pedagogical gestures illustrating at least one L2 phonetic or prosodic feature. Second, the training effects should be measured in at least one modality (perception and/or production) by either comparing performance before and after the training (i.e., within-subjects design) or comparing at least two groups of participants (i.e., between-subjects design). We carefully screened studies published in major journals in L2 research (e.g., *Language Learning*, *Studies in Second Language Acquisition*, *Journal of Second Language Pronunciation*) and extended our search to conference proceedings (e.g., *Gesture and Speech in Interaction*) to identify potential sources. Google Scholar was used as an additional source to check for any relevant studies. Finally, we identified 24 empirical studies involving pedagogical gestures in L2 phonetic training. Given the purposes identified above, we focused on three main factors for each study: training paradigm (observing gestures vs. performing gestures¹), testing modality (perception vs. production), and training effects (immediate gains, generalization, and retention). Other details, such as participants' demographic information, are summarized in Appendix A.

2. Beat Gestures

Beat gestures, characterized by up and down hand movements in sync with speech prominence (McNeill, 1992), are closely associated with prosodically prominent parts of speech and can act as visual markers of speech prominence (Loehr, 2012). When speakers perform a visual beat on an emphasized word, listeners tend to perceive that word as more acoustically prominent (Krahmer & Swerts, 2007). In L2 speech, beat gestures can be used to signal speech prominence, which may aid L2 learners' speech comprehension and production (Mccafferty, 2006). These benefits are grounded in the deeper biomechanical and cognitive coupling between gestures and speech. For instance, the hand gesture apex tends to align with pitch-accent peaks (Esteve-Gibert & Prieto, 2013), the maximum gesture velocity coincides with emphasized syllables (Danner et al., 2018), and finger tapping interacts with speech prominence (Parrell et al., 2014). Beat gestures can directly modulate speech acoustics through respiration-related mechanisms. When speakers perform beat gestures, their pitch and amplitude peaks systematically occur near the moment of the maximum arm extension (Pouw et al., 2020a, 2021). In short, beat gestures actively contribute to the temporal and prosodic structure of speech (Prieto et al., 2025a).

Given the link between beat gestures and speech prominence, empirical research has explored the application of beat gestures in L2 phonetic training, which has yielded positive results. Gluhareva and Prieto (2017) investigated how observing beat gestures affects L2 pronunciation accuracy at the discourse level. In a within-subjects design, Catalan-Spanish bilinguals practiced English pronunciation by watching pre-recorded videos. Half of the training materials featured beat gestures highlighting speech prominence, and the other half did

not. No generalization or retention effects were examined. The results showed that observing beat gestures could reduce learners' accentedness in their oral responses to difficult discourse scenarios. In a between-subjects design, [Llanes-Coromina et al. \(2018\)](#) assessed whether performing beat gestures improved Catalan-Spanish bilinguals' English oral proficiency better than no gesture training. The results showed that actively performing beat gestures led to better global oral proficiency (accentedness, comprehensibility, and fluency) than no gesture training. Crucially, the reading materials differed across the pretest, training, and posttest, which means that performing beat gestures yielded generalization effects. Although beat gestures benefited L2 oral proficiency, it was unclear whether performing beat gestures was better than merely observing them. To address this gap, [Prieto et al. \(2025b\)](#) assessed whether the training paradigm influences the effectiveness of beat gestures. Using the same contexts as [Gluhareva and Prieto \(2017\)](#), two groups of Catalan-Spanish bilinguals were trained either by performing or by observing beat gestures. The results showed that actively performing beat gestures while verbally repeating the training sentences yielded more reduced accentedness in English from pretest to posttest than merely observing the gestures. However, no generalization effects were observed in either group. The results suggest that training paradigms play an important role in the effectiveness of gestural training. Taken together, it seems that at the discourse level, beat gestures can improve learners' global oral proficiency, and performing gestures leads to more gains than merely observing gestures. However, the robustness of beat gesture training was unclear, as generalization effects were not consistently confirmed, and retention effects were not tested in any of the three studies.

Beyond the scope of discourse-level prominence, some empirical studies have attempted to use beat gestures on other phonetic aspects, such as lexical stress and vowel length, though findings have been inconsistent. [Hirata et al. \(2014\)](#) trained English speakers to identify Japanese long and short vowels using beat gestures. In their design, two beats represented a long vowel, and one beat represented a short vowel². One group of participants only observed the gestures, whereas the other group performed the gestures. Both groups showed similar amounts of improvement in the perceptual accuracy of the Japanese vowel length contrast from pretest to posttest, suggesting that the efficacy of beat gestures is not modulated by the training paradigm. In addition, no generalization effects were found in this experiment, as reported in [Kelly et al. \(2014\)](#). However, as no control group was included in the design, it was not possible to assess the effects of gestural training *per se*. [van Maastricht et al. \(2019\)](#) trained Dutch speakers to learn Spanish lexical stress under one of three conditions. In the metaphorical gesture condition, learners observed hand gestures mimicking the enhanced duration of stressed syllables, where the instructor moved both hands to the side of the body. In the beat gesture condition, learners observed beat gesture strokes on the stressed syllables. In the control condition, no gestures were shown. The improvement from pretest to posttest was assessed by the accuracy of lexical stress placement in speech production elicited from a sentence reading task. Notably, the training only featured words. Therefore, this design essentially allowed testing generalization effects in a new context, that is, from the word level to the sentence level. However, all three groups of participants showed similar gains, which suggests that neither type of gesture facilitated learning lexical stress production. Consequently, no generalization effects could be concluded.

Beat gestures seem to boost L2 phonetic training more effectively at the discourse level than at the word and segment levels, especially in speech production. Therefore, the robustness of beat gestures in L2 learning may be more pronounced in signaling discourse-level speech prominence, a domain that is more closely related to global L2 oral proficiency. Moreover, the generalization and retention effects remain largely unexplored. These gaps suggest potential directions for future studies. Table 1 summarizes the studies reviewed in this subsection (see "beat" and "beat & durational" sections) and Table A1 in Appendix A provides more detailed information about the experimental design.

Table 1. Summary of empirical studies testing the effects of gestural training on L2 speech learning.

Author (Year)	Target	Paradigm	Modality	Task	Measure	Effects		
						Immediate After Training	Generalization	Retention at Delayed Test After Training
Beat								
Gluhareva and Prieto (2017)	Speech prominence	Obs.	Prod.	Discourse completion	Perceptual rating (accentedness)	+	NA	NA
Llanes-Coromina et al. (2018)	Speech prominence	Perf.	Prod.	Text reading	Perceptual rating (accentedness, comprehensibility, and fluency)	+	+(untrained item)	NA
Prieto et al. (2025b)	Speech prominence	Obs. vs. Perf.	Prod.	Discourse completion	Perceptual rating (accentedness)	+(Perf. > Obs.)	× (untrained item)	NA
Beat & Durational								
Hirata and Kelly (2010)	Vowel length	Obs.	Perc.	Identification	Accuracy rate	×	× (new context)	NA
Hirata et al. (2014)	Vowel length	Obs. vs. Perf.	Perc.	Identification	Accuracy rate	×	NA	NA
Kelly et al. (2014)	Vowel length	Obs. vs. Perf.	Perc.	Identification	Accuracy rate	×	× (untrained item)	NA
van Maastricht et al. (2019)	Stress	Obs.	Prod.	Sentence reading	Accuracy (on-target vs. non-target)	×	×	NA
Durational								
Li et al. (2020)	Vowel length	Perf.	Perc. & Prod.	Identification & word imitation	Accuracy rate & vowel duration ratio	+(Prod. only)	× (untrained item)	NA
Shimada (2025)	Vowel length	Perf.	Perc. & Prod.	Identification & word reading	Accuracy rate & comprehensibility rating	+	+(untrained item)	+(3 weeks)
Pitch								
Morett and Chang (2015)	Lexical tone	Perf.	Perc.	Word tone identification & word meaning association	Accuracy rate	+	× (untrained item)	NA
Baills et al. (2019)	Lexical tone	Obs. vs. Perf.	Perc.	Tone identification & word meaning recall & word meaning association	Accuracy rate	+(Obs. = Perf.)	× (untrained item)	NA
Zhen et al. (2019)	Lexical tone	Obs. vs. Perf.	Perc.	Identification	Accuracy rate	+(Perf. > Obs. only when gestures are performed in horizontal plane)	+(untrained item; Perf. > Obs. only when gestures are performed in horizontal plane)	+(1 day; Perf. > Obs. only when gestures are performed in horizontal plane)

Table 1. Cont.

Author (Year)	Target	Paradigm	Modality	Task	Measure	Effects		
						Immediate After Training	Generalization	Retention at Delayed Test After Training
Yu et al. (2024)	Lexical tone	Perf.	Perc.	Discrimination & word-picture association	Accuracy rate	+	NA	NA
Gao et al. (2026)	Tone sandhi	Perf.	Prod.	Word reading	Perceptual rating	×	+(untrained item)	NA
Hirata et al. (2024)	Pitch accent	Perf.	Perc.	Identification	Accuracy rate	×	NA	NA
Shimada (2025)	Pitch accent	Perf.	Perc. & Prod.	Identification & word reading	Accuracy rate & comprehensibility rating	+	+	+(3 weeks)
Yuan et al. (2019)	Nuclear accent	Obs.	Prod.	Discourse completion	Accuracy (on-target vs. non-target)	+	NA	NA
Baills et al. (2022)	Intonation	Obs.	Prod.	Dialogue reading	Perceptual rating (accentedness, comprehensibility, fluency, suprasegmental accuracy, and segmental accuracy)	+(accentedness & suprasegmental accuracy)	× (untrained item)	NA
Li et al. (2023b)	Intonation	Obs.	Prod.	Dialogue reading & sentence imitation	Perceptual rating (accentedness, comprehensibility, fluency for dialogue-reading and accentedness for sentence-imitation)	+(accentedness)	× (untrained item)	+(2 weeks)
Gestures illustrating articulatory features								
Amand and Touhami (2016)	Unreleased stops	Obs.	Prod.	Word & sentence reading	Accuracy (on-target vs. non-target)	+	NA	NA
Amand and Touhami (2024)	Unreleased stops	Obs.	Prod.	Word & sentence reading	Accuracy (on-target vs. non-target)	+	NA	+(1 month)
Hoetjes and van Maastricht (2020)	/u/ and /θ/	Obs.	Prod.	Sentence reading	Accuracy (on-target vs. non-target) & perceptual rating (accentedness, comprehensibility)	Mixed (Iconic for /u/; Pointing for both)	Mixed (new context, same as immediate gains)	NA
Xi et al. (2020)	Aspirated stops and affricates	Obs.	Perc. & Prod.	Identification & word imitation	Accuracy rate & perceptual rating (aspiration accuracy and overall pronunciation)	+(Stops Prod. only)	NA	NA

Table 1. Cont.

Author (Year)	Target	Paradigm	Modality	Task	Measure	Effects		
						Immediate After Training	Generalization	Retention at Delayed Test After Training
Li et al. (2021)	Aspirated stops	Perf.	Perc. & Prod.	Identification & word imitation	Accuracy rate & perceptual rating (overall pronunciation) and voice onset time	+(Prod. only, accurate gesture performance only)	× (untrained item)	+(3 days, Prod. only)
Xi et al. (2024)	/æ-ʌ/	Obs.	Perc. & Prod.	Identification & word imitation, text reading, picture naming	Accuracy rate & formant analysis	+(Prod. only, visible articulatory movement only)	× (untrained item)	+(1 week, Prod. only)
Gao et al. (2026)	/u/ and aspirated stops	Perf.	Prod.	word reading	Mahalanobis distance & voice onset time	+	+(untrained item)	NA

Note. Perc. = speech perception; Prod. = speech production; Obs. = learners observe hand gestures without enacting them; Perf. = learners perform hand gestures after viewing the instructor’s gestures; + = positive effect; × = null effect; > = better than; = = equal to; NA = not available.

3. Durational Gestures

In some languages, duration is phonemic and distinguishes word meanings. For example, in languages such as Japanese, Italian, and Finnish, vowels or consonants contrast in length, which may pose challenges for L2 learners whose L1 does not have such length distinctions. One proposed approach to facilitating the acquisition of these contrasts involves using hand gestures that metaphorically represent duration. This mapping is supported by behavioral evidence. Casasanto and Boroditsky (2008) showed that spatial movement strongly influences individuals' estimations of temporal duration. Building on this, Cai and Connell (2012) demonstrated that when spatial length is perceived through both tactile and visual modalities, it greatly affects the estimation of temporal duration. Furthermore, Cai et al. (2013) showed that hand gestures moving in space can significantly shape listeners' temporal judgments. These behavioral studies imply that phonological durational contrasts can be effectively represented by contrasting horizontal hand movements, which are called durational gestures in later literature.

However, previous research has generally reported null effects of durational gestures on the perceptual learning of L2 length contrasts, either by observing or by performing such gestures. Hirata and Kelly (2010) trained English speakers to identify Japanese vowel length contrasts. The hand gestures involved in this study consisted of a beat gesture showing short vowels and a durational gesture representing long vowels. Participants were trained under one of four conditions. In the audio-only condition, participants could only rely on auditory input to practice the perception of vowel length. In the audio-hand condition, participants had access to hand gestures but could not see the speaker's mouth. In the audio-mouth condition, no gesture was provided, but participants could see the speaker's mouth. In the audio-mouth-hand condition, participants had access to both the mouth and hand gestures. Participants only observed the gestures without performing them. Perceptual learning was compared between a pretest and a posttest via a vowel length identification task. Although the audio-mouth condition significantly improved learners' vowel length perception, observing hand gestures during training did not show any extra benefit. Although the training used nonsense syllables and the test materials were real words, the lack of gestural benefits ruled out the possibility of observing generalization effects. In a follow-up study with a similar design, Hirata et al. (2014) compared two gesture types and two training paradigms. The first type, syllable gestures, used a single beat for short vowels and a durational gesture for long vowels. The second type, mora gestures, used one beat for short vowels and two beats for long vowels. The English speakers either observed or performed these gestures during training, yielding a 2 gesture type $\times$ 2 training paradigm design. The four training conditions led to similar improvements in the perception of the Japanese vowel length contrast, although observing syllable gestures appeared to yield more balanced improvements when the target sounds varied by word position or speech rate. The generalization of training effects to untrained items, as well as potential impacts on vocabulary learning, were examined in Kelly et al. (2014). However, all four training conditions showed similar generalization effects, and no significant group differences were observed in the vocabulary learning outcomes. These null results from behavioral experiments were further supported by electrophysiological evidence from Kelly and Hirata (2017), who showed that gestures did not modulate neural responses to vowel length distinctions during early stages of processing. Taken together, the series of studies conducted by Hirata, Kelly, and colleagues suggests that neither observing nor performing durational gestures and/or beat gestures can facilitate the perceptual learning of L2 vowel length. As all these studies only featured a pretest and a posttest, it was not possible to identify retention effects.

Nevertheless, durational gestures have been proven effective in improving the production of L2 length contrasts. Instead of using beat gestures, which typically highlight prominence and may not be suitable for cueing short vowels due to their lower visual salience compared to long vowels, Li et al. (2020) adapted the gesture design for vowel length contrasts by using horizontal hand movements for both short and long vowels. Short vowels were paired with short spatial movements, and long vowels were paired with long spatial movements. Two groups of Catalan-Spanish bilinguals were trained on both the perception and production of the Japanese vowel length contrast. One group repeated the instructor's speech without any gesture input, and the other repeated the speech while performing gestures. Results showed that the gesture group significantly improved their vowel length production accuracy, as measured by the duration ratio of long and short vowels, from pretest to posttest, whereas the no gesture group did not. By contrast, no significant between-group difference was found in perceptual improvement, which aligns with prior findings. In addition, gestural training did not show generalization effects in perception or production, and retention effects were not assessed. These lab studies (Hirata & Kelly, 2010; Hirata et al., 2014; Li et al., 2020; Kelly et al., 2014) revealed the unbalanced effects of durational and beat gestures in the perception and production of L2 vowel length, which suggests a potential dissociation between the two modalities in the phoneme-level phonetic learning.

Beyond lab settings, durational gestures have been introduced into classroom teaching, which has yielded positive effects. In a multi-session training design with a pretest/posttest/delayed posttest paradigm, Shimada (2025) trained English speakers to learn the Japanese vowel length contrast. One group of participants received training by performing durational gestures as described by Li et al. (2020), while the other group underwent computer-assisted training. Learners' perception was assessed using an identification task, and production was assessed using a word reading task, which was later rated for comprehensibility. Results showed that gestural training led to significant gains in both perception and production of vowel length, and these improvements were retained after three weeks. Moreover, the training effects were generalized to new words. Although gestural training was overall comparable to computer-assisted training, learners found the gestures to be intuitive and helpful. However, without a no-training control group, it is difficult to determine the specific contribution of the gestures. Therefore, the positive effects reported by Shimada (2025) cannot overturn the null findings for durational gestures in the perceptual modality, as reported in earlier studies. In addition, comprehensibility ratings may not target the fine-grained phonetic aspects of vowel length contrasts, which should be complemented by an acoustic analysis.

In summary, while the theoretical grounding for mapping temporal duration to spatial hand movement is robust, empirical evidence suggests that durational gestures offer limited benefits for L2 perceptual learning regarding phonological length contrasts, though they may support production gains. Future research should explore refined gesture designs and multimodal training methods that can improve both perception and production. Moreover, given the relatively small number of studies on durational gestures, many research questions remain unexplored. First, it is unclear whether and how performing gestures is more effective than observing gestures in speech production. Second, more evidence is needed to validate generalization and retention effects. These gaps suggest potential future directions for gestural studies. The studies reviewed in this section are summarized in the "beat & durational" and "durational" sections of Table 1. More detailed information on the experimental design is available in Table A1 of Appendix A.

4. Pitch Gestures

Pitch plays an important role in the perception and production of speech at both the word level (e.g., lexical tones, stress, and pitch accent) and sentence level (e.g., intonation). Pitch features can be trained with gestural cues because the perception of acoustic pitch and hand movements in space share similar representational and processing resources. Casasanto et al. (2003) presented participants with lines moving vertically from bottom to top or horizontally from left to right and asked them to reproduce either the displacement or pitch of auditory stimuli. The results indicated that vertical displacement strongly influenced participants' estimates of acoustic pitch, whereas horizontal displacement did not. This supports the existence of a conceptual metaphor linking vertical space and pitch height. Connell et al. (2013) further demonstrated that watching upward or downward gestures biases listeners' perception of acoustic pitch as being higher or lower than its actual pitch height, supporting the notion of a shared representation of pitch and spatial height. Neurophysiological evidence also supports this link, as judging auditory pitch height activates the unimodal visual areas of the brain, indicating an overlap in the processing of auditory pitch and visuospatial height (Dolscheid et al., 2014). Given these findings, there is a clear potential for integrating hand movements in the vertical plane to metaphorically represent pitch in speech during the training of L2 pitch features.

Pitch gestures can significantly improve the perception of pitch variations, with the majority of studies focusing on L2 lexical tones. In a seminal study, Morett and Chang (2015) taught monosyllabic Mandarin words to three groups of English speakers. In the pitch gesture condition, participants performed hand gestures which depicted the pitch contours of Mandarin lexical tones. In the semantic gesture condition, participants performed hand gestures that iconically depicted word meanings. In the no gesture condition, participants received training without gestures. Participants' learning outcomes were assessed by a tone identification task and a word-meaning association task in the pretest and posttest. The results showed that pitch gestures facilitated participants' memorization of Mandarin words contrasting in lexical tones, which suggests that pitch gestures can enhance the association between lexical meaning and tones. However, no generalization effects were observed in gestural training. Baills et al. (2019) extended Morett and Chang's (2015) line of research by comparing the effects of observing and performing pitch gestures. They trained Catalan-Spanish bilinguals to learn Mandarin tones in monosyllabic words by either observing or performing pitch gestures. Both observing and performing pitch gestures showed better results than no gestural training in L2 lexical tone perception, as measured by a tone identification task and a word-meaning association task in the pretest and posttest. However, performing gestures did not show more benefits than observing them. Again, no generalization effects were observed. Later, Zhen et al. (2019) showed that the direction in which pitch gestures are performed is crucial to the perceptual learning of L2 lexical tones. Specifically, observing and performing pitch gestures performed in the vertical plane (e.g., moving the hand from bottom to top for a rising tone) resulted in similar training effects, replicating Baills et al.'s (2019) findings. However, when gestures were performed in the horizontal plane (e.g., moving the hand away from the body to indicate a rising tone), L1 English participants benefited more from performing gestures than from merely observing them in terms of L2 Mandarin lexical tone perception. Similar patterns were revealed in a generalization test with new items and a follow-up test, suggesting training retention effects, although the time interval between the posttest and follow-up test was relatively short (only one day). Additionally, the benefits of pitch gestures on perceptual learning of L2 lexical tones can be extended to L1 tone language speakers. Yu et al. (2024) trained Mandarin speakers in the perception of Thai tones under one of three conditions: (a) observing diagrams depicting pitch contours, (b) performing gestures while observing

pitch contour diagrams, and (c) word-picture association without gestures. The results showed that performing hand gestures yielded the best results for perceptual learning of L2 lexical tones by comparing the participants' tone identification and word-picture association accuracy from pretest to posttest. Although the tests included untrained words, the authors did not report an analysis of generalization effects.

Based on the training studies reviewed thus far, we can start resolving the puzzles regarding how the training paradigm (observing gestures vs. performing gestures) affects learning outcomes. In the context of L2 lexical tone perception, the effectiveness of this paradigm appears to depend on the nature of the gesture itself. Specifically, when gestures are presented in a canonical way, observing gestures yields training effects comparable to those of performing gestures. In the case of pitch gestures, the vertical plane allows for straightforward metaphorical mapping between pitch and spatial hand movements (i.e., high pitch corresponds to "high" in space, and low pitch corresponds to "low" in space). Under such conditions, observing vertically performed pitch gestures is as effective as performing them. However, when gestures are presented in a non-canonical manner, that is, when the gesture does not align with the conventional metaphorical relationship between the phonetic concept and spatial representation (e.g., pitch gestures performed in the horizontal plane), observing gestures appears to be less beneficial than performing them. These findings highlight the importance of considering how gestures are performed when comparing the effects of performing gestures versus observing gestures in L2 phonetic training. We will return to this point in Section 5.

Unlike perceptual learning, evidence on the role of pitch gestures in L2 lexical tone production is scarce and mixed. Gao et al. (2026) conducted a multi-session classroom training study in which Japanese speakers learned Mandarin tone sandhi, where Tone 3 (dipping pitch) becomes Tone 2 (rising pitch) before another Tone 3 by performing pitch gestures following the instructor. Learners' production elicited from a word reading task was rated by native listeners for tone sandhi accuracy. Although gestures did not significantly improve learners' tone sandhi production, as assessed by comparing the trained items in the pretest against the posttest, gestural training yielded better generalization to new items. Retention effects were not tested. These results suggest that pitch gestures may facilitate the transfer of tonal knowledge beyond trained contexts. However, this single study is not sufficient to draw firm conclusions regarding the benefits of gestures in L2 lexical tone production. This suggests a potential research gap for future studies, where more attention should be paid to speech production.

Apart from lexical tones, a small number of studies have examined the use of pitch gestures to train L2 lexical stress and lexical pitch accent, with mixed results. Positive results were found in a classroom study. Shimada (2025) trained English speakers to perceive and produce Japanese pitch accent by performing pitch gestures showing accent contrasts. Participants showed significant gains in both perception and production from pretest to posttest, maintained these gains after three weeks, and generalized their knowledge from trained to untrained items. By contrast, a recent lab study showed limited effects. Hirata et al. (2024) compared auditory-only, visual pitch-height notation, and pitch gestures for English speakers learning Japanese pitch accent. Pitch-height notation led to the greatest perceptual improvements, which were generalized to untrained words. However, performing pitch gestures did not enhance learning. A follow-up experiment with left- and right-hand pitch gestures also found no perceptual advantage for gestural training. Overall, in learning word stress and pitch accent, Shimada's (2025) classroom study has revealed more benefits of pitch gestures than Hirata et al.'s lab study. It might be that learners need multiple sessions and a long training period to consolidate their knowledge. This calls

for future studies to conduct longitudinal and classroom studies to assess the long-lasting effects and ecological validity of gestural training.

Pitch gestures are also beneficial for the perception and production of larger linguistic units, such as intonation. Yuan et al. (2019) trained two groups of Mandarin speakers to practice Spanish sentences intonation. One group observed gestures illustrating nuclear contours and the other group received audio-only training. Learners' production was assessed through a discourse completion task and later analyzed for nuclear pitch accent accuracy. Observing pitch gestures yielded greater improvement than no gesture training from pretest to posttest, suggesting that pitch gestures can help the production of L2 intonation. Moving beyond nuclear accents, Baills et al. (2022) demonstrated that observing pitch gestures accompanying logatome (i.e., "dadada" non-sense syllable sequences that mimicked intonation contours) enhanced Catalan-Spanish bilinguals' accuracy in L2 French suprasegmentals and reduced their accentedness, as assessed by a dialogue-reading task in the pretest and posttest. Li et al. (2023b) showed that observing pitch gestures tracing intonation contours of entire sentences reduced Catalan-Spanish bilinguals' L2 French accentedness, with gains retained at a delayed posttest conducted two weeks after training. Among all three studies, only Yuan et al. (2019) found generalization effects from trained to untrained items, whereas Baills et al. (2022) and Li et al. (2023b) did not. In addition, more evidence is needed to assess how pitch gestures help in the perception of L2 intonation, which can be a future research line.

In summary, as the most widely investigated pedagogical gesture, pitch gestures have been shown to be beneficial for learning pitch-related features in an L2. The most robust evidence comes from L2 lexical tone perception and intonation production, while the results for lexical stress and pitch accent are mixed. A possible reason is that pitch gestures are most effective when the target prosodic feature involves dynamic contour movement, such as tone or intonation, allowing gestures to map transparently onto the auditory pattern. In contrast, when pitch information is less contour-based and more categorical, such as stress or pitch accent, gestures may provide less representational support. Further controlled, contrastive research is needed to test this hypothesis and clarify the conditions under which pitch gestures are effective across prosodic domains. All the papers reviewed in this section are summarized in the "pitch" section of Table 1. For the details of the experimental design, see Table A1 in Appendix A.

5. Hand Gestures Illustrating Articulatory Features

Hand gestures can be used for phonetic training on the segmental level because they can iconically represent the articulatory movements or features of L2 vowels and consonants, or they can draw learners' attention to the target articulatory features by pointing to articulatory movements. In these teaching scenarios, again, learners can either observe and/or perform the instructor's pedagogical gestures to enhance their understanding of the target articulatory features. Human's arm movements and speech act are governed by a shared sensorimotor control system, which leads to mutual influence between arm movements and speech production (Gentilucci & Corballis, 2006; Gentilucci & Dalla Volta, 2008; Pouw et al., 2020b, 2020c, 2021). For instance, when speakers observe a hand grasping motion while articulating vowels and consonants, their lip aperture and speech voice amplitude increase if the grasping action indicates a larger object (Gentilucci, 2003). Similarly, when speakers produce non-sense syllables while grasping objects, the size of the objects can also affect the mouth aperture and other articulatory gestures relevant to the target sounds (Gentilucci et al., 2001). In other words, the shape of hand gestures may directly affect articulatory gestures, which can support L2 pronunciation training.

Building on the concept of hand–mouth coupling, a series of studies have developed different pedagogical gestures to correspond with a variety of target phonemes. [Amand and Touhami \(2016\)](#) conducted a between-subjects study to examine whether gestures could aid French speakers in correctly pronouncing English word-final stops, which are often unreleased in English but less so in French. In their study, they designed two gestures to show released and unreleased plosive stops. For released stops, the instructor curled her fingers into a fist and then abruptly opened her hand, spreading all five fingers outward. This rapid outward motion metaphorically illustrates the burst of air typical of a released or aspirated stop. For unreleased stops, the instructor closed five fingers into a firm fist without any outward release, indicating that no audible airburst should occur with unreleased stops. In the gesture condition, learners could perform the gestures. Learners' production of English stops was assessed through a sentence and word reading task administered in a pretest and posttest, which were later acoustically analyzed for accuracy. The results showed that while participants generally improved in producing unreleased stops, those trained with hand gestures showed significantly greater improvement than those without gestures. Taken together, these results demonstrate that gestures can boost learners' awareness of aspiration patterns. A follow-up study by [Amand and Touhami \(2024\)](#) replicated these results, and further showed that learners retained knowledge after one month using the same gestural training. However, the retention effects observed in [Amand and Touhami \(2024\)](#) should be interpreted with caution, as only six ($n = 3$ per group) out of the 28 posttest participants returned to the delayed posttest, which questions the statistical power. Moreover, in both studies, the gesture condition combined two multimodal cues: hand gestures and tactile cues. The latter involved placing the finger in front of the mouth to feel the aspirated airflow. This design therefore made it difficult to isolate the role of gestures. In light of Amand and Touhami's gestures for stops, subsequent research has systematically explored the conditions under which hand gestures effectively facilitate L2 sound acquisition, with the most important factor being the link between hand gestures and the target phonetic features to be learned. The following factors appear to influence the training effects.

First, pedagogical gestures should iconically or metaphorically represent the target L2 features. [Xi et al. \(2020\)](#) adapted the fist-to-open-palm gesture from [Amand and Touhami \(2016\)](#) for released stops to train word-initial aspirated stops. Two groups of Catalan speakers were trained to learn six Mandarin consonants, with or without observing hand gestures. Three of these pairs were plosives /p-p^h, t-t^h, k-k^h/, distinguished by aspiration and the absence/presence of a strong air burst. The remaining three pairs were affricates /ts-ts^h, tʃ-tʃ^h, tʂ-tʂ^h/, which differ acoustically in the duration of the frication period. Learners' perception was tested through an identification task and production through a word-imitation task, which was later rated for production accuracy. The results revealed that while observing the fist-to-open-palm gesture significantly improved participants' pronunciation of plosives from pretest to posttest, it failed to improve the pronunciation of affricates. This may be attributed to the gesture's mimicry of a strong airburst for aspiration, which might be inadequate to cue the durational features pertinent to the affricate pairs. In addition, gestural training did not show much benefit for perceptual learning of the target Mandarin consonants. In addition, no generalization effects were observed. [Hoetjes and van Maastricht \(2020\)](#) examined how observing pointing gestures and iconic gestures influenced Dutch speakers' learning of the L2 Spanish sounds /u/ and /θ/. Participants' production was tested with a sentence reading task, which was later rated by native speakers. The results revealed that pointing to the articulatory features of the target sounds showed similar benefits for the production of /u/ and /θ/. However, iconic gestures showed sound-specific effects. The gesture for /u/, which involved rounding

the palm to depict lip rounding, supported production learning, whereas the gesture for /θ/, which used extended fingers to represent tongue protrusion, impeded production learning. Notably, since the training only featured exemplar words, the gains reported here actually signified a generalization effect from the word level to the sentence level. These findings highlight an interesting interaction between the complexity of gesture shape and phoneme-to-be-learned. That is, gestures used to train L2 phonemes should be adequate; when the phoneme is relatively difficult, overly complex gestures may impose additional cognitive demands and thus hinder learning. Finally, no generalization or retention effects were reported.

Second, the efficacy of gestural training is conditional on learners' ability to accurately imitate the instructor's gestures. Li et al. (2021) trained Catalan-Spanish bilinguals to learn Mandarin aspirated plosives with or without hand gestures using the same type of gestures as in Xi et al. (2020). Participants learned novel words containing the target plosives and were tested on perception and production through a pretest, immediate posttest, and delayed posttest. The tasks were identical to those of Xi et al. (2020). The results showed that learners who appropriately performed the fist-to-open-palm gesture significantly improved their production accuracy compared to those who either struggled with performing gestures or received non-gestural training. More importantly, accurate gesture performance triggered better retention effects, as shown by the delayed posttests administered three days after training, although no generalization effects were found. Li et al. (2024) analyzed the gesture group of Li et al. (2021) and found that during the training phase, the accuracy of learners' production of aspirated stop consonants was influenced by their gesture performance quality. The authors identified two critical predictors of gesture performance related to speech production: the correctness of the gesture shape and temporal alignment between speech and gesture. Specifically, learners who maintained a firm fist for an adequate duration and those who opened their fists prior to speech onset showed more accurate pronunciation of the target aspirated consonants. These results together suggest that learners' gesture performance quality is a key predictor of successful gestural training in L2 sound acquisition. In practice, if instructors encourage learners to actively perform pedagogical gestures, they should ensure that learners are capable of accurately performing the gestures.

Third, the visibility of the articulatory movements encoded by hand gestures appears to be an important factor influencing the training effects. Gestures encoding visible articulatory movements provide a stronger boost than those representing less visible or non-visible movements. Hoetjes and van Maastricht's (2020) findings on Dutch learners learning Spanish /u/ and /θ/ can also be interpreted from this angle. Specifically, for /u/, the rounding-palm gesture for *visible* lip rounding facilitated learning, whereas the extended-finger gesture for the *less-visible* tongue protrusion hindered learning of /θ/. However, because the learning difficulty of the two phonemes is also related to orthography, the results cannot be taken as direct evidence for the visibility-based account proposed here. Building on this, Xi et al. (2024) compared the effects of gestural training on visible versus non-visible articulatory features on the learning of English vowels /æ/ and /ʌ/ by Catalan-Spanish bilinguals. These vowels differ in *visible* lip aperture and *non-visible* tongue position. Participants received training either (a) with no gestures, (b) with observing a gesture illustrating lip opening, or (c) with observing a gesture depicting tongue position. The participants were tested in an identification task for perception; a word-imitation task, a text-reading task, and a picture-naming task for production. Although gestures did not show much benefit in perceptual learning, gestures highlighting the visible lip aperture led to greater improvements in vowel production, as revealed by acoustic analysis. The "lip hand gesture" also showed retention effects after one week, but no generalization effects

were found. Together, these findings indicate that pedagogical gestures are most effective when they map onto the articulatory properties that learners can visually perceive and readily interpret.

Despite previous studies showing robust training effects on the trained items, and some also showing retention effects after training, it is not clear whether gains obtained through gestural training can be generalized to untrained items in speech production. To address this gap, Gao et al. (2026) adopted the gesture for /u/ (Hoetjes & van Maastricht, 2020) and the gesture for aspirated stops (Li et al., 2021; Xi et al., 2020) in a classroom study with a multi-session training program. The participants were L1 Japanese-speaking learners of Mandarin. Mandarin /u/ and aspirated stops are challenging for the learners. Japanese /u/ is less rounded than Mandarin /u/; and Japanese voiceless stops are less aspirated than those in Mandarin. Participants' speech production was elicited through a word-reading task and later acoustically analyzed for its accuracy. The results showed that training with performing hand gestures could help learners improve their pronunciation accuracy from pretest to posttest and generalize their knowledge to untrained items. This study thus contributes to a comprehensive understanding of how gestures illustrating articulatory features can support L2 pronunciation. As previous studies have consistently shown that gestures illustrating articulatory features do not help perception, Gao et al. (2026) did not implement a perception test. This can be a future topic for classroom studies.

To summarize, the empirical evidence regarding the role of pedagogical gestures illustrating articulatory features expands on previous research on prosodic learning. First, this collection of studies confirms the limited role of gestures in perceptual learning at the word- and phoneme-level, but reveals robust training effects in improving production. This points to the asymmetrical developmental trajectories between L2 perception and production in the context of multimodal phonetic training. Second, the mediating factors that constrained the positive effects of performing gestures over observing gestures were clarified, which warrants classroom studies on similar topics. Finally, the robustness of gestural training has been evaluated beyond the immediate gains which includes generalization and retention effects, although, results on generalization effects are less consistent than those of retention effects. The "gestures illustrating articulatory features" section in Table 1 summarizes all the studies reviewed in this section. See Table A1 in Appendix A for details of the experimental design.

6. Discussion

In this review, we synthesized the effectiveness of four types of pedagogical gestures (i.e., beat gesture, durational gesture, pitch gesture, and gestures illustrating articulatory features) on L2 phonetic training. The studies show sufficient variability in terms of the learners' L1 backgrounds and the target L2 phonetic and prosodic features. Our main conclusions are as follows.

First, the effectiveness of hand gestures on L2 speech acquisition differs by modality, with more robust training effects observed in production than in perception. A considerable number of studies have either not assessed the effectiveness of hand gestures on L2 perceptual learning or have failed to identify positive effects, particularly concerning durational and segmental features (see Table 1 for a summary). By contrast, gestures have shown positive effects on the perceptual learning of suprasegmental features, such as lexical tones (Baills et al., 2019; Morett & Chang, 2015; Yu et al., 2024). Several reasons might account for this inconsistency. For one, speech perception relies on multiple acoustic cues, but hand gestures typically focus on a single acoustic dimension (e.g., lip aperture for the /æ-ʌ/ contrast). Given individual differences in perceptual cue-weighting (Chandrasekaran et al., 2010; Kong, 2019; Kong & Kang, 2023), the specific acoustic domain illustrated by gestural

training may not align with the primary cue that L2 learners use to distinguish difficult sound pairs. Moreover, unlike suprasegmental features such as intonation, segmental features span a single segment within a limited timeframe. Therefore, during perceptual events, learners may not have sufficient time to process multimodal information, potentially affecting learning outcomes (Hirata et al., 2024). Given that perception and production do not always develop concurrently in L2 speech acquisition (Nagle & Baese-Berk, 2022), it is not surprising that measurable gains appear in one modality (e.g., production) but not necessarily in the other (e.g., perception).

Second, the positive role of hand gestures seems to depend on various mediating factors, among which gesture-sound consistency and the visibility of the target L2 feature are the most widely validated. Gesture-sound consistency includes (a) whether the hand gesture adequately represents the learning target iconically or metaphorically (Hirata et al., 2014; Hirata & Kelly, 2010; Kelly et al., 2014; Xi et al., 2020), and (b) whether the learners' gesture performance is on-target (Li et al., 2021). This implies that pedagogical gestures are not merely visual reminders but meaningful symbols referring to a specific phonetic or prosodic feature. The lack of positive effects of beat gestures on the perceptual learning of L2 vowel length (Hirata & Kelly, 2010; Hirata et al., 2014; Kelly et al., 2014) may also be attributed to the fact that beat gestures may not match the phonetic features of duration. If this assumption holds, hand gestures should outperform meaningless visual cues as well. Furthermore, the scope of the target features matters. Beat gestures appear robust for discourse-level prominence but less so for word-level features such as vowel length (e.g., Hirata & Kelly, 2010; Hirata et al., 2014; Kelly et al., 2014) or lexical stress (van Maastricht et al., 2019). This means that gesture-sound consistency must be evaluated at an adequate linguistic level. Regarding visibility, the visual saliency of the target L2 phonetic/prosodic feature largely affects the training effects. The best results were obtained from visually salient features (e.g., mouth aperture, Xi et al., 2024), followed by visually less salient features (e.g., interdental tongue position like /θ/, Hoetjes & van Maastricht, 2020), and least for non-visible features (e.g., tongue position inside the mouth, Xi et al., 2024). When gestures attempt to encode non-visible articulatory features, such as tongue shape or position, they may fail to provide clear iconic mapping and instead increase the cognitive load. In other words, when the articulatory target is hidden, complex hand movements may distract rather than guide the learner.

Third, the robustness of a training paradigm can be assessed through immediate training effects, retention, and generalization. Most studies found positive effects of hand gestures by comparing the pretest to the posttest immediately after the training intervention. A subset of studies administered delayed posttests, confirming that gestural benefits can be retained over time (Amand & Touhami, 2024; Li et al., 2021, 2023b; Shimada, 2025; Xi et al., 2024). Moreover, although the evidence is mixed, two recent studies (Gao et al., 2026; Shimada, 2025) demonstrated successful generalization effects in multi-session classroom teaching. This points to the fact that the null results of generalization effects found in lab studies might be due to insufficient training sessions. These findings suggest that gestural training has the potential to help learners internalize and retain their knowledge acquired during multimodal phonetic training interventions, although in practice, more training sessions might further favor generalization effects.

Despite the positive findings, some major research gaps emerge in this review, which can be summarized into five aspects. Briefly, the role of observing gestures and performing gestures is not clear; hand gestures do not consistently benefit L2 perception and production; ecological validity and longitudinal evidence are needed; individual differences appear to interact with training effects; and the neural mechanisms underlying the beneficial role of

hand gestures have yet to be thoroughly explored. In what follows, we elaborate on these five major gaps.

Observing gestures versus performing gestures: While embodied cognition theories propose that motor engagement can enhance memory retention and facilitate learning processes, empirical findings have been inconsistent. For instance, [Prieto et al. \(2025b\)](#) and [Llanes-Coromina et al. \(2018\)](#) found that learners who actively performed hand gestures achieved significantly greater improvements in global L2 oral proficiency compared to those who merely observed gestures. However, other research (e.g., [Baills et al., 2019](#)) reported that both observing and performing gestures were equally effective for learning Mandarin tones. Furthermore, [Zhen et al. \(2019\)](#) introduced a spatial nuance, which emphasized that the advantage of performing gestures could only be found when the gestures were observed/performed in the horizontal plane. If the gestures were observed/performed in the vertical plane, observing or performing gestures would yield similar beneficial results. These results suggest that the advantage of performing gestures may be modulated by the gesture type, spatial direction, or specific linguistic feature being targeted. This line of research merits future empirical studies to advance our understanding of the effective integration of hand gestures into L2 speech learning.

Relationship between speech perception and production: The effectiveness of gestures appears to be stronger in L2 speech production than in perception. This disconnect implies that perception may not necessarily precede production in L2 speech acquisition and that gestures may influence these two modalities via distinct cognitive pathways. Early studies on L2 speech acquisition revealed mixed results regarding whether perception and production correlate with each other (see [Flege & Bohn, 2021](#), for a comprehensive discussion). Recent studies suggest that the two modalities may have distinct developmental patterns ([Nagle & Baese-Berk, 2022](#)) and may co-evolve without precedence ([Flege & Bohn, 2021](#)). Therefore, to thoroughly understand this dynamic relationship, researchers should conduct longitudinal observations. Consistent with this idea, training studies often involve multiple testing sessions. For instance, some studies reviewed in this article adopted a pretest/posttest/delayed posttest design (e.g., [Li et al., 2023b](#); [Xi et al., 2024](#)). This can offer an opportunity to observe how the perception-production relationship evolves given the proper intervention. Exploring this research question within the scope of multimodal phonetic training can complement the findings of other non-intervention studies and deepen our understanding of the theoretical aspects of L2 speech acquisition.

Ecological validity and longitudinal effects: The majority of reviewed articles reported strictly controlled lab-based studies, with little longitudinal evidence. Many studies recruited novice learners to exclude the potential influence of learners' pre-existing knowledge of the target language on training effects. Only a few studies have been conducted in classroom training contexts. However, some lack rigorous control conditions, which limits the generalizability of their findings (e.g., [Shimada, 2025](#)). In addition, retention effects have either not been tested or assessed only after short intervals (e.g., 1 week). Finally, only part of the studies found generalization effects, others either did not test this effect or failed to find significant generalizability of gestural training. More importantly, to facilitate analytical measures, most studies used controlled speech production tasks, such as imitation and reading tasks, with limited tests on spontaneous speech. Consequently, it is not clear whether the immediate gains observed in the lab can translate to long-term L2 proficiency or spontaneous speech in real-world communicative contexts. These limitations question the ecological validity and longitudinal effects of multimodal phonetic training in practice, which calls for future studies to explore this perspective.

Individual differences: The success of gestural training is likely mediated by individual learner differences, a factor often overlooked in group-level analyses. [Li et al. \(2021\)](#),

2024) highlighted that the quality of gesture performance is a significant predictor of the learning outcomes. This implies that motoric ability or “gesture aptitude,” may be a potential individual factor to consider in future work. If a learner struggles to coordinate hand and mouth movements, multimodal cues may present an additional cognitive challenge, which potentially hinders rather than helps L2 speech acquisition. Beyond gesture-related aptitude, some studies have included other cognitive factors, such as musical abilities (e.g., Li et al., 2020; Yuan et al., 2019) and working memory (e.g., Li et al., 2020). However, systemic evaluations of how individual cognitive differences interact with gestural training are needed.

Neural mechanisms: The neural mechanisms underlying the role of gestures in L2 phonetic learning remain unclear. Most research has assessed the role of hand gestures using behavioral measures, with limited neurophysiological evidence (but see Kelly & Hirata, 2017). Understanding the neural underpinnings of how hand gestures modulate auditory processing and speech learning is crucial for refining pedagogical interventions.

7. Conclusions and Future Directions

This review synthesizes over a decade of empirical research on the role of hand gestures in L2 phonetic training. The literature reviewed indicates that the integration of visual and motoric cues encoded by hand gestures reduces cognitive demands, which allows learners to process abstract phonological concepts more effectively. Existing research indicates that pedagogical gestures (i.e., beat gestures, durational gestures, pitch gestures, and gestures illustrating articulatory features) can significantly enhance L2 speech acquisition. Specifically, hand gestures show the most robust effects when they are congruent with the target phonetic or prosodic features (e.g., pitch gestures for tone) and when they align with visible articulatory properties (e.g., lip shape).

However, the effectiveness of gestural training is not uniform. As illustrated in this review, there are significant gaps concerning the optimal paradigm of training (observing gestures vs. performing gestures), the influence of individual learner differences, and the long-term retention of these skills in ecological classroom settings. Furthermore, the dissociation often found between perceptual and productive gains challenges our understanding of the sensorimotor mechanisms involved. Future research must move beyond “does it work?” to “how and for whom does it work?” by incorporating neuroimaging techniques and longitudinal designs to unravel the neural mechanisms underlying embodied L2 learning. Bridging these gaps will facilitate the development of more effective multimodal pedagogies that make full use of the body–mind connection.

Author Contributions: Conceptualization, X.X. and P.L.; methodology, X.X. and P.L.; writing—original draft preparation, P.L.; writing—review and editing, X.X. and P.L.; funding acquisition, X.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Social Science Fund of China, grant number 24CYY082.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

L2 Second language

L1 First language

Appendix A. Additional Information About Experimental Design of the Empirical Studies Involved in the Review

Table A1. Demographic information and experimental design.

Author (Year)	Age ^a	AoA & YoL	L1-L2 Pair	L2 Proficiency	Training Context	Training Conditions & Sample Size
Beat						
Gluhareva and Prieto (2017)	19.3 (3.31)	NA	CA/ES-EN	Upper Int.	lab	20 (within-subjects)
Llanes-Coromina et al. (2018)	14.08 (0.28)	NA	CA/ES-EN	Ele.-lower Int.	lab	59 in total (beat vs. no beat)
Prieto et al. (2025b)	21.5 (3.32)	NA	CA/ES-EN	Upper Int. (B2)	lab	9 gesture observation, 9 gesture production
Beat & Durational						
Hirata and Kelly (2010)	[19, 23]	NA	EN-JA	Nov.	lab	15 audio-only, 15 audio-mouth, 16 audio-hands, 14 audio-mouth-hands
Hirata et al. (2014)	[18, 23]	NA	EN-JA	Nov.	lab	22 each (syllable gesture observation, syllable gesture production, mora gesture observation, mora gesture production)
Kelly et al. (2014)	[18, 23]	NA	EN-JA	Nov.	lab	22 each (syllable gesture observation, syllable gesture production, mora gesture observation, mora gesture production)
van Maastricht et al. (2019)	25 [18, 65]	NA	NL-ES	Nov.	lab	62 in total (no gesture, beat gesture, metaphoric gesture)
Durational						
Li et al. (2020)	19.86 [18, 29]	NA	CA/ES-JA	Nov.	lab	25 each (no gesture vs. gesture)
Shimada (2025)	NA	NA	EN-JA	Ele.-Int.	classroom	9 gesture vs. 8 computer-assisted
Pitch						
Morett and Chang (2015)	20.33 (3.34)	NA	EN-ZH	Nov.	lab	57 in total (no gesture, pitch gesture, semantic gesture)
Baills et al. (2019)	Exp1: 19.86 (1.44) Exp2: 19.93 (1.41)	NA	CA/ES-FR	Nov.	lab	Exp1: 24 no gesture, 25 gesture observation; Exp2: 28 no gesture production vs. 28 gesture production)
Zhen et al. (2019)	21.65 (2.52)	NA	EN-ZH	Nov.	lab	18 each (auditory only, congruent pitch gesture production, congruent pitch gesture observation, rotated pitch gesture production, rotated pitch gesture observation, incongruent pitch gesture production)
Yu et al. (2024)	[18, 23]	NA	ZH-TH	Nov.	lab	30 each (word-picture association, pitch gesture production, pitch feature observation)
Gao et al. (2026)	20.46 (1.52)	AoA = 18 YoL = 3	JA-ZH	Ele.-Int.	classroom	20 gesture, 19 no gesture
Hirata et al. (2024)	[17, 22]	NA	EN-JA	Nov.	lab	66 in total (flat notation, notation, gesture)
Shimada (2025)	NA	NA	EN-JA	Ele.-Int.	classroom	9 gesture vs. 8 computer-assisted
Yuan et al. (2019)	19.80 (1.30)	YoL = 0.18	ZH-ES	Beginner (A1)	lab	32 each (no gesture vs. gesture)

Table A1. Cont.

Author (Year)	Age ^a	AoA & YoL	L1-L2 Pair	L2 Proficiency	Training Context	Training Conditions & Sample Size
Baills et al. (2022)	19.88 (2.01) ^b	YoL = 4.58 ^b	CA/ES-FR	Int. (A2-B1)	lab	27 speech only, 22 non-embodied logatome, 26 embodied logatome
Li et al. (2023b)	19.89 (3.63)	AoA = 13.72 YoL = 5.52	CA/ES-FR	Ele.-Int.	lab	28 embodied vs. 29 non-embodied
Gestures illustrating articulatory features						
Amand and Touhami (2016)	NA	NA	FR-EN	Ele.-Adv.	lab	8 each (gesture vs. no gesture)
Amand and Touhami (2024)	NA	NA	FR-EN	NA	lab	Pretest: control = 20, gesture = 42; training & posttest: control = 17, gesture = 14; delayed: control = 3, gesture = 3
Hoetjes and van Maastricht (2020)	25 [18, 61]	NA	NL-ES	Nov.	lab	12 audio-only, 13 audiovisual, 13 pointing gesture, 12 iconic gesture
Xi et al. (2020)	20.90 (2.50)	NA	CA/ES-ZH	Nov.	lab	25 each (no gesture vs. gesture)
Li et al. (2021)	19.31 (1.64)	NA	CA/ES-ZH	Nov.	lab	29 no gesture, 29 gesture, 9 control
Xi et al. (2024)	19.7 (1.8)	YoL = 5.17 ^b	CA/ES-ZH	NA	lab	33 each (no gesture, lip hand gesture, tongue hand gesture)
Gao et al. (2026)	20.46 (1.52)	AoA = 18 YoL = 3	JA-ZH	Ele.-Int.	classroom	20 gesture, 19 no gesture

^a. Age information is not consistently reported across studies. Depending on the available numeric information, we cite the mean age, the standard deviation (in parentheses), or the age range [lower limit, upper limit]. ^b. The authors calculated the mean from the reported data. Note: AoA = age of acquisition; YoL = years of formal learning; CA = Catalan; EN = English; ES = Spanish; FR = French; JA = Japanese; NL = Dutch; TH = Thai; ZH = Chinese; Ele. = elementary; Int. = intermediate; Adv. = advanced; Nov. = novice (i.e., without prior knowledge in the target L2); NA = not available.

Notes

- ¹ In the literature, if learners actively perform the pedagogical gestures, the training paradigm is often called “gesture production” or “producing gestures.” To avoid confusion with speech production, in this review, we use the verb “perform.” Nevertheless, in Table 1, we retained the original condition names to be consistent with the cited literature.
- ² Hirata et al. (2014) refer to this movement as a “mora gesture,” which they distinguish from a “syllable gesture.” The latter consists of a beat gesture representing a short vowel and a horizontal hand movement representing a long vowel. As the hand sweep for a long vowel belongs to the scope of “durational gestures,” we discuss this study more thoroughly in Section 3.

References

- Allen, L. Q. (1995). The effects of emblematic gestures on the development and access of mental representations of French expressions. *The Modern Language Journal*, 79(4), 521–529. [\[CrossRef\]](#)
- Amand, M., & Touhami, Z. (2016). Teaching the pronunciation of sentence final and word boundary stops to French learners of English: Distracted imitation versus audio-visual explanations. *Research in Language*, 14(4), 377–388. [\[CrossRef\]](#)
- Amand, M., & Touhami, Z. (2024). Could you say [læpˈtɒp]? Acquisition of unreleased stops by advanced French learners of English using spectrograms and gestures. *Languages*, 9(8), 257. [\[CrossRef\]](#)
- Baills, F., Alazard-Guiou, C., & Prieto, P. (2022). Embodied prosodic training helps improve accentedness and suprasegmental accuracy. *Applied Linguistics*, 43(4), 776–804. [\[CrossRef\]](#)
- Baills, F., Suárez-González, N., González-Fuente, S., & Prieto, P. (2019). Observing and producing pitch gestures facilitates the learning of Mandarin Chinese tones and words. *Studies in Second Language Acquisition*, 41(1), 33–58. [\[CrossRef\]](#)
- Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, 59, 617–645. [\[CrossRef\]](#)
- Barsalou, L. W. (2010). Grounded cognition: Past, present, and future. *Topics in Cognitive Science*, 2(4), 716–724. [\[CrossRef\]](#)
- Cai, Z. G., & Connell, L. (2012). Space-time interdependence and sensory modalities: Time affects space in the hand but not in the eye. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th annual conference of the cognitive science society* (pp. 168–173). Cognitive Science Society.
- Cai, Z. G., Connell, L., & Holler, J. (2013). Time does not flow without language: Spatial distance affects temporal duration regardless of movement or direction. *Psychonomic Bulletin and Review*, 20(5), 973–980. [\[CrossRef\]](#)

- Casasanto, D., & Boroditsky, L. (2008). Time in the mind: Using space to think about time. *Cognition*, 106(2), 579–593. [[CrossRef](#)] [[PubMed](#)]
- 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. Kirsh (Eds.), *Proceedings of the annual meeting of the cognitive science society* (Vol. 25, p. 1323). Cognitive Science Society.
- Chandrasekaran, B., Sampath, P. D., & Wong, P. C. M. (2010). Individual variability in cue-weighting and lexical tone learning. *The Journal of the Acoustical Society of America*, 128(1), 456–465. [[CrossRef](#)]
- Connell, L., Cai, Z. G., & Holler, J. (2013). Do you see what I'm singing? Visuospatial movement biases pitch perception. *Brain and Cognition*, 81, 124–130. [[CrossRef](#)]
- Danner, S. G., Barbosa, A. V., & Goldstein, L. (2018). Quantitative analysis of multimodal speech data. *Journal of Phonetics*, 71, 268–283. [[CrossRef](#)]
- Dolscheid, S., Willems, R. M., Hagoort, P., & Casasanto, D. (2014). The relation of space and musical pitch in the brain. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *The 36th annual meeting of the cognitive science society* (Vol. 36, Issue 2014, pp. 421–426). California Digital Library, University of California.
- 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. [[CrossRef](#)]
- Flege, J. E., & Bohn, O.-S. (2021). The revised speech learning model (SLM-r). In R. Wayland (Ed.), *Second language speech learning: Theoretical and empirical progress* (pp. 3–83). Cambridge University Press.
- Gao, S., Xi, X., & Li, P. (2026). Revisiting the benefits of hand gestures in L2 pronunciation: Generalization effects in multi-session multimodal phonetic training. *Language and Speech*. [[CrossRef](#)] [[PubMed](#)]
- Gentilucci, M. (2003). Grasp observation influences speech production. *European Journal of Neuroscience*, 17(1), 179–184. [[CrossRef](#)] [[PubMed](#)]
- Gentilucci, M., Benuzzi, F., Gangitano, M., & Grimaldi, S. (2001). Grasp with hand and mouth: A kinematic study on healthy subjects. *Journal of Neurophysiology*, 86(4), 1685–1699. [[CrossRef](#)]
- Gentilucci, M., & Corballis, M. C. (2006). From manual gesture to speech: A gradual transition. *Neuroscience and Biobehavioral Reviews*, 30(7), 949–960. [[CrossRef](#)] [[PubMed](#)]
- 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. [[CrossRef](#)]
- Gluhareva, D., & Prieto, P. (2017). Training with rhythmic beat gestures benefits L2 pronunciation in discourse-demanding situations. *Language Teaching Research*, 21(5), 609–631. [[CrossRef](#)]
- Gullberg, M. (2022). The relationship between gestures and speaking in L2 learning. In T. M. Derwing, M. J. Munro, & R. I. Thomson (Eds.), *The Routledge handbook of second language acquisition and speaking* (pp. 386–398). Routledge. [[CrossRef](#)]
- Hirata, Y., Friedman, E., Kaicher, C., & Kelly, S. D. (2024). Multimodal training on L2 Japanese pitch accent: Learning outcomes, neural correlates and subjective assessments. *Language and Cognition*, 16(4), 1718–1755. [[CrossRef](#)]
- 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(2), 298–310. [[CrossRef](#)]
- 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. [[CrossRef](#)]
- Hoetjes, M., & van Maastricht, L. (2020). Using gesture to facilitate L2 phoneme acquisition: The importance of gesture and phoneme complexity. *Frontiers in Psychology*, 11, 575032. [[CrossRef](#)] [[PubMed](#)]
- Ionescu, T., & Vasc, D. (2014). Embodied cognition: Challenges for psychology and education. *Procedia-Social and Behavioral Sciences*, 128, 275–280. [[CrossRef](#)]
- Iverson, J. M., & Goldin-Meadow, S. (2005). Gesture paves the way for language development. *Psychological Science*, 16(5), 367–371. [[CrossRef](#)]
- Kelly, S. D., & Hirata, Y. (2017). What neural measures reveal about foreign language learning of Japanese vowel length contrasts with hand gestures. In S. Tanaka, G. Pinter, S. Ogawa, M. Giriko, & H. Takeyasu (Eds.), *New development in phonology research: Festschrift in honor of haruo kubozone* (pp. 278–294). Kaitakusha.
- 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. [[CrossRef](#)] [[PubMed](#)]
- Kelly, S. D., & Lee, A. L. (2012). When actions speak too much louder than words: Hand gestures disrupt word learning when phonetic demands are high. *Language and Cognitive Processes*, 27(6), 793–807. [[CrossRef](#)]
- Kelly, S. D., McDevitt, T., & Esch, M. (2009). Brief training with co-speech gesture lends a hand to word learning in a foreign language. *Language and Cognitive Processes*, 24(2), 313–334. [[CrossRef](#)]
- Kong, E. J. (2019). Individual differences in categorical perception: L1 English learners' L2 perception of Korean stops. *Phonetics and Speech Sciences*, 11(4), 63–70. [[CrossRef](#)]

- Kong, E. J., & Kang, S. (2023). Individual differences in categorical judgment of L2 stops: A link to proficiency and acoustic cue-weighting. *Language and Speech*, 66(2), 354–380. [\[CrossRef\]](#)
- Kontra, C., Lyons, D. J., Fischer, S. M., & Beilock, S. L. (2015). Physical experience enhances science learning. *Psychological Science*, 26(6), 737–749. [\[CrossRef\]](#)
- 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(3), 396–414. [\[CrossRef\]](#)
- Li, P., Baills, F., Alazard-Guiu, C., Baqué, L., & Prieto, P. (2023a). A pedagogical note on teaching L2 prosody and speech sounds using hand gestures. *Journal of Second Language Pronunciation*, 9(3), 340–349. [\[CrossRef\]](#)
- Li, P., Baills, F., Baqué, L., & Prieto, P. (2023b). The effectiveness of embodied prosodic training in L2 accentedness and vowel accuracy. *Second Language Research*, 39(4), 1077–1105. [\[CrossRef\]](#)
- 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. [\[CrossRef\]](#)
- Li, P., Baills, F., Xi, X., & Prieto, P. (2024). Gesture shape and gesture-speech alignment predict simultaneous L2 sound production accuracy. In A. Brown, & S. W. Eskildsen (Eds.), *Multimodality across epistemologies in second language research* (pp. 173–186). Routledge. [\[CrossRef\]](#)
- 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. [\[CrossRef\]](#)
- Llanes-Coromina, J., Prieto, P., & Rohrer, P. L. (2018). Brief training with rhythmic beat gestures helps L2 pronunciation in a reading aloud task. In K. Klessa, J. Bachan, A. Wagner, M. Karpiński, & D. Śledziński (Eds.), *Proceedings of the international conference on speech prosody* (pp. 498–502). ISCA. [\[CrossRef\]](#)
- Loehr, D. P. (2012). Temporal, structural, and pragmatic synchrony between intonation and gesture. *Laboratory Phonology*, 3(1), 71–89. [\[CrossRef\]](#)
- Logan, J. S., & Pruitt, J. S. (1995). Methodological issues in training listeners to perceive non-native phonemes. In W. Strange (Ed.), *Speech perception & linguistic experience: Issues in cross-language* (pp. 351–378). York Press.
- Macedonia, M. (2014). Bringing back the body into the mind: Gestures enhance word learning in foreign language. *Frontiers in Psychology*, 5, 1467. [\[CrossRef\]](#)
- Matheson, H. E., & Barsalou, L. W. (2018). Embodiment and grounding in cognitive neuroscience. In J. Wixted (Ed.), *Stevens' handbook of experimental psychology and cognitive neuroscience* (pp. 357–383). John Wiley & Sons, Inc.
- Mccafferty, S. G. (2006). Gesture and the materialization of second language prosody. *IRAL-International Review of Applied Linguistics in Language Teaching*, 44(2), 197–209. [\[CrossRef\]](#)
- McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago Press.
- Mizelle, J. C., & Wheaton, L. A. (2010). Why is that hammer in my coffee? A multimodal imaging investigation of contextually based tool understanding. *Frontiers in Human Neuroscience* 4, 233. [\[CrossRef\]](#)
- Morett, L. M., & Chang, L. Y. (2015). Emphasising sound and meaning: Pitch gestures enhance Mandarin lexical tone acquisition. *Language, Cognition and Neuroscience*, 30(3), 347–353. [\[CrossRef\]](#)
- Nagle, C., & Baese-Berk, M. M. (2022). Advancing the state of the art in L2 speech perception-production research: Revisiting theoretical assumptions and methodological practices. *Studies in Second Language Acquisition*, 44(2), 580–605. [\[CrossRef\]](#)
- Parrell, B., Goldstein, L., Lee, S., & Byrd, D. (2014). Spatiotemporal coupling between speech and manual motor actions. *Journal of Phonetics*, 42, 1–11. [\[CrossRef\]](#) [\[PubMed\]](#)
- 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. [\[CrossRef\]](#) [\[PubMed\]](#)
- Pouw, W., Harrison, S. J., & Dixon, J. A. (2020a). Gesture-speech physics: The biomechanical basis for the emergence of gesture-speech synchrony. *Journal of Experimental Psychology: General*, 149(2), 391–404. [\[CrossRef\]](#)
- Pouw, W., Paxton, A., Harrison, S. J., & Dixon, J. A. (2020b). Acoustic information about upper limb movement in voicing. *Proceedings of the National Academy of Sciences of the United States of America*, 117(21), 11364–11367. [\[CrossRef\]](#)
- Pouw, W., Trujillo, J. P., & Dixon, J. A. (2020c). The quantification of gesture-speech synchrony: A tutorial and validation of multimodal data acquisition using device-based and video-based motion tracking. *Behavior Research Methods*, 52(2), 723–740. [\[CrossRef\]](#)
- Prieto, P., Esteve-Gibert, N., & Shattuck-Hufnagel, S. (2025a). Towards a novel conceptualization of prosody that accounts for spoken and visual signals: The modality-neutral prosodic framework hypothesis. *Gesture*, 23(1–2), 119–159. [\[CrossRef\]](#)
- Prieto, P., Kushch, O., Borràs-Comes, J., Gluhareva, D., & Pérez-Vidal, C. (2025b). Training ESL students to reproduce beat gestures in discourse leads to L2 pronunciation improvements. *Anuario Del Seminario de Filología Vasca "Julio de Urquijo"*, 57(1–2), 805–823. [\[CrossRef\]](#)
- Rato, A., & Oliveira, D. (2023). Assessing the robustness of L2 perceptual training: A closer look at generalization and retention of learning. In U. Kickhöfel Alves, & J. I. Alcantara de Albuquerque (Eds.), *Assessing the robustness of L2 perceptual training: A closer look at generalization and retention of learning* (pp. 369–396). De Gruyter Mouton. [\[CrossRef\]](#)

- Shapiro, L., & Stolz, S. A. (2019). Embodied cognition and its significance for education. *Theory and Research in Education*, 17(1), 19–39. [\[CrossRef\]](#)
- Shimada, M. (2025). A comparison of techniques for training L2 Japanese prosody. *Journal of Second Language Pronunciation*, 11(2), 148–173. [\[CrossRef\]](#)
- Sweller, J., van Merriënboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, 10(3), 251–296. [\[CrossRef\]](#)
- Tellier, M. (2008). The effect of gestures on second language memorisation by young children. *Gesture*, 8(2), 219–235. [\[CrossRef\]](#)
- van Maastricht, L., Hoetjes, M., & van Drie, E. (2019). Do gestures during training facilitate L2 lexical stress acquisition by Dutch learners of Spanish? In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th international congress of phonetic sciences* (pp. 6–10). Australasian Speech Science and Technology Association Inc. [\[CrossRef\]](#)
- Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin and Review*, 9(4), 625–636. [\[CrossRef\]](#)
- 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. [\[CrossRef\]](#)
- Xi, X., Li, P., & Prieto, P. (2024). Improving second language vowel production with hand gestures encoding visible articulation: Evidence from picture-naming and paragraph-reading tasks. *Language Learning*, 74(4), 884–916. [\[CrossRef\]](#)
- Yang, J., & Shu, H. (2016). Involvement of the motor system in comprehension of non-literal action language: A meta-analysis study. *Brain Topography*, 29(1), 94–107. [\[CrossRef\]](#)
- Yu, K., Zhang, J., Li, Z., Zhang, X., Cai, H., Li, L., & Wang, R. (2024). Production rather than observation: Comparison between the roles of embodiment and conceptual metaphor in L2 lexical tone learning. *Learning and Instruction*, 92, 101905. [\[CrossRef\]](#)
- 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. [\[CrossRef\]](#)
- Zhen, A., van Hedger, S., Heald, S., Goldin-Meadow, S., & Tian, X. (2019). Manual directional gestures facilitate cross-modal perceptual learning. *Cognition*, 187, 178–187. [\[CrossRef\]](#) [\[PubMed\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.