LANGUAGE LEARNING
A Journal of Research in Language Studies
Language Learning ISSN 0023-8333

EMPIRICAL STUDY

Improving Second Language Vowel Production With Hand Gestures Encoding Visible Articulation: Evidence From Picture-Naming and Paragraph-Reading Tasks

Xiaotong Xi [a,b] Peng Li [c] and Pilar Prieto [d,a]

[a]Universitat Pompeu Fabra [b]Shandong University of Finance and Economics [c]University of Oslo [d]Institució Catalana de Recerca i Estudis Avançats

CRediT author statement – **Xiaotong Xi**: conceptualization (lead); methodology (equal); writing – original draft (lead); writing – review and editing (equal); formal analysis (lead). **Peng Li**: conceptualization (supporting); methodology (equal); writing – review and editing (equal); formal analysis (supporting). **Pilar Prieto**: conceptualization (supporting); methodology (supporting); writing – review and editing (equal); supervision (lead).

A one-page Accessible Summary of this article in nontechnical language is freely available in the Supporting Information online and at https://oasis-database.org

The study was supported by the Ministerio de Ciencia, Innovación y Universidades, Agencia Estatal de Investigación, and Fondo Europeo de Desarrollo Regional [PGC2018-097007-B-I00], and by the Ministerio de Ciencia e Innovación [PID2021-123823NB-I00]. The first author is supported by the Department d'Empresa i Coneixement de la Generalitat de Catalunya and the European Social Fund under a grant for the recruitment of early-stage research staff [2021FI_B 00137]. The second author is supported by the Research Council of Norway through its Centres of Excellence funding scheme [223265].

Many thanks to the English teachers at Universitat Pompeu Fabra, Almudena Diaz, Aodh O'Byrne, Carmen Pérez-Vidal, Geòrgia Pujadas, Judith Borràs, Kayla Chapin, Laura Pons Tortosa, Mary Lofgren, Patrick Louis Rohrer, and Sharon Gutiérrez Metelli, for their help in recruiting participants for this experiment. We particularly thank Patrick Louis Rohrer for helping to create the materials. Finally, we thank all the participants for volunteering to take part in this study.

Correspondence concerning this article should be addressed to Xiaotong Xi, Shandong University of Finance and Economics, 40 Shungeng Road, Jinan, Shandong Province, China. Email: xiaotong.xi@sdufe.edu.cn

The handling editor for this manuscript was Sarah Grey.

DOI: 10.1111/lang.12647

Abstract: This study investigates whether audiovisual phonetic training with hand gestures encoding visible or nonvisible articulation features has a differential impact on learning second language sounds. Ninety-nine Catalan–Spanish bilingual students were trained to differentiate English /æ/ and /ʌ/, which differ in the visible lip aperture and nonvisible tongue position, with training involving no gestures, gestures representing the lip aperture, or gestures representing the tongue position. Before, immediately after, and 1 week after the training, participants' perception of the targets was assessed through a word-identification task, and their production was tested through paragraph-reading, picture-naming, and word-imitation tasks. Although all participants improved in perception and production, the lip hand gesture was more effective in adjusting lip aperture than the other two conditions in the paragraph-reading and picture-naming tasks. These results suggest that hand gestures encoding visible rather than nonvisible articulation features are more effective for improving second language pronunciation.

Keywords hand gesture; phonetic training; English; vowel; second language acquisition

Introduction

In the domain of second language (L2) acquisition of pronunciation, adult learners often experience difficulties in perceiving and producing L2 sounds. While perception-based phonetic training typically yields improvements in both areas, researchers have begun to explore the effectiveness of integrating hand gestures encoding phonetic information into such training. However, this type of training has not consistently demonstrated benefits. Therefore, this study aims to investigate whether the visibility of the articulatory features encoded by hand gestures used in the training plays a crucial role in its effectiveness. Specifically, we will compare the effects of training with hand gestures mimicking visible lip aperture, training with hand gestures mimicking nonvisible tongue position, and training without hand gestures on the learning of L2 English vowel contrast /æ/-/ʌ/ by Catalan–Spanish bilinguals.

Background Literature

Perception-Based Phonetic Training

Research in L2 acquisition has revealed that acquiring L2 sounds that are not present in learners' first language can present challenges due to many variables, including limited exposure to the L2 (e.g., Muñoz, 2008) and the quality of instruction (e.g., Derwing & Munro, 2015). These challenges often manifest as L2 adult learners struggling to accurately perceive and produce certain sounds of the target language. Perception-based phonetic training provides learners

with necessary L2 input, thereby helping learners improve their perception and production of nonnative sounds.

In its canonical form, traditional perception-based training involves audio input only (for review of high variability phonetic training, see Barriuso & Hayes-Harb, 2018), but researchers have attempted to add the visual modality to provide access to the speaker's face during the phonetic training sessions. Audiovisual phonetic training led to greater improvements in L2 perception and production compared to audio-only high variability phonetic training (Hardison, 2003; Hazan et al., 2005; Inceoglu, 2016; for null results, see Reyes & Hazan, 2021).

Interestingly, Hazan et al. (2005) revealed that audiovisual phonetic training was more effective than auditory-only phonetic training in improving perception performance when the articulation difference between the contrasted L2 sounds was visually salient (e.g., bilabial /b/ vs. labiodental /v/) compared to a less visually salient contrast (e.g., lateral approximant /l/ vs. central approximant /r/). Thus, the benefits obtained from integrating auditory and visual input in the learning of L2 sound contrasts seem to depend upon the *visual accessibility* of the target articulation difference. When the articulation difference is visually not accessible, the mere sight of the face may not provide sufficient phonetic information to favor L2 learning. Providing the visual modality for the tongue, which is hardly visible in its full motion and configurations, may allow L2 learners to access an important part of articulatory information.

However, in the realm of speech processing, previous studies have observed the null effects of seeing the tongue. For example, viewing the computer-animated tongue movement did not lead to higher identification accuracy of speech sounds than not seeing it (Grauwinkel et al., 2007). Similarly, with normal auditory input, exposure to two visual modalities of the visible mouth and the animated tongue at the same time did not lead to higher identification accuracy than merely observing the mouth (Wik & Engwall, 2008). This suggests that interpreting intra-oral articulation during speech perception can be difficult, as humans are not accustomed to seeing this type of information in daily communication. In addition, when both visual modalities are provided, listeners attend more to the mouth than the tongue (Badin et al., 2010), which may suggest the redundant role of viewing the tongue in speech processing. Despite these null effects, computer-assisted phonetic training incorporates a talking head simulating articulatory movements. Learners can either observe the talking head's tongue movement or observe their own articulation during sound production, which has been shown to facilitate the perception and production of L2 sounds more strongly than training

without the simulated articulation (Pillot-Loiseau et al., 2013; Wang et al., 2014).

Apart from using computer-assisted techniques, another layer of visual supporting information can be provided by hand gestures. Hand gestures are frequently used by language teachers in L2 classrooms to visually highlight phonological features of L2 speech and promote pronunciation learning (Hudson, 2011; Smotrova, 2017). In the following subsection, we will summarize the effects of hand gestures on learning L2 speech sounds.

Audiovisual Phonetic Training With Hand Gestures

Audiovisual phonetic training with hand gestures involves multiple information resources, including auditory input of speech sounds, visible articulatory information from the instructor's face (such as lips, teeth, and tongue tip), and visual input of hand gestures encoding phonetic features. To investigate the potential benefits of hand gestures in training L2 sounds, researchers typically compare the learning outcomes of an audiovisual phonetic training paradigm with hand gestures to those of an audiovisual phonetic training condition without hand gestures. However, the findings from previous studies have produced mixed results.

At the perception level, training involving hand gestures encoding acoustic information such as vowel duration or air burst did not achieve greater gains than training without gestures (Hirata & Kelly, 2010; Hirata et al., 2014; Li et al., 2020, 2021; Xi et al., 2020). Concerning L2 production, the positive effects of hand gestures were not systematic. For example, hand gestures encoding durational features facilitated the production of L2 Japanese long and short vowels (Li et al., 2020). Similarly, hand gestures mimicking aspiration revealed positive effects for learning L2 plosives (English released and unreleased stops: Amand & Touhami, 2016; Mandarin aspirated and unaspirated plosives: Li et al., 2021; Xi et al., 2020).

However, mixed effects have been observed when hand gestures encode articulatory information. Hoetjes and van Maastricht (2020) found that using a hand gesture forming an "o" shape to mimic lip roundness for L2 Spanish /u/ was helpful for Dutch speakers' production of the target sound, but this proved not to be the case for another gesture in which the fingers were extended forward to mimic the tongue protruding for the Spanish /θ/. One possible explanation for the null effect of the latter gesture may have been its form. As Dutch speakers tend to mispronounce the Spanish interdental fricative /θ/ as alveolar fricatives /s/ or /z/, perhaps a gesture focusing on the positioning of the tongue between the teeth rather than merely its shape would have proved

more effective. In addition, it seems reasonable to conjecture that hand gestures mimicking more visually accessible external articulatory cues (lip roundness for the Spanish /u/) will be more effective than those mimicking less visible ones (tongue protrusion for the Spanish /θ/). However, the authors noted that there was a distinction in learning difficulty between the two sounds. Spanish /θ/ is a nonnative sound for Dutch speakers and is more complex compared to Spanish /u/, which is present in the first language Dutch phonetic inventory although the phoneme-to-grapheme conversion differs across these two languages. Taking this into consideration, the comparison of the learning outcomes between /u/ and /θ/ is not sufficient to confirm the conjecture of the visibility effect.

Though the issue clearly requires deeper exploration, there are to our knowledge only two studies involving the use of hand gesture mimicking tongue shape within the mouth (i.e., when it is not visible) to teach pronunciation. In the first, a classroom study, the instructor bent or flattened the fingers to mimic the tongue shapes required for English /r/ and /l/ respectively (Lan & Wu, 2013), which helped students to improve the pronunciation accuracy of the two target sounds. In the second study, from a clinical context, an adult patient with apraxia was trained to pronounce the /r/ sound (Rusiewicz & Rivera, 2017). The patient's pronunciation of the /r/ sound improved after receiving treatment with hand gestures mimicking the tongue shape of the target sound.

Considering the potential role played by the visibility of articulators encoded by hand gestures in training L2 sounds (Hoetjes & van Maastricht, 2020), the main goal of the present study is to assess the potential differential effects of hand gestures encoding visible versus nonvisible articulatory features (specifically, features involving lips vs. tongue) on the learning of L2 sounds.

The positive effects of including hand gestures in audiovisual phonetic training are supported by several theoretical frameworks. According to the dual coding theory, information coded through both verbal and nonverbal channels is easier to retain and retrieve compared to that conveyed through unimodal channels (Clark & Paivio, 1991). Thus, in the event that the mental representation from the verbal channel is degraded or forgotten, another extra layer of representation from the nonverbal channel (e.g., hand gestures) can play a compensatory role in retrieving information from memory and thus boost learning (Paivio, 1991). From the perspective of embodied cognition theory, physical actions can shape cognitive processes (e.g., Wilson, 2002). Since watching another person's behavior can activate the neurons that prime the execution of the same action (called mirror neurons; Rizzolatti & Craighero, 2004), either viewing teachers performing hand gestures or self-performing gestures

enhances students' learning outcomes (for a review of empirical evidence, see Sullivan, 2018). Following up on these claims, observing hand gestures should trigger stronger learning effects than not seeing them during audiovisual phonetic training.

The Present Study

The present study investigates whether the effects of audiovisual phonetic training with hand gestures might be influenced by the visibility of the articulatory features encoded by those gestures. We selected the target vowel pair /æ/–/ʌ/, as these vowels differ in both the degree of lip aperture (and thus in vowel height), which is visible, and tongue position (i.e., vowel backness), which is nonvisible. Regarding vowel height, the low-front vowel /æ/ is produced with more jaw opening than the mid-back vowel /ʌ/ (Fromkin, 1964). As for vowel backness, the tongue dorsum is more fronted in /æ/ than in /ʌ/ (Blackwood Ximenes et al., 2017). These articulatory features are directly related to acoustic features, specifically to the frequencies of the first formant (F1) and second formant (F2): The low-front /æ/ shows higher frequencies for these two formants (F1: 660–1,010 Hz; F2: 1,720–2,320 Hz) than does the mid-back /ʌ/ (F1: 640–850 Hz; F2: 1,190–1,590 Hz; Peterson & Barney, 1952).

The target population was Catalan–Spanish bilinguals learning English. Since neither Catalan nor Spanish have a comparable vowel contrast in the low central region of the vowel space, Catalan–Spanish bilinguals learning English tend to perceptually assimilate the two English vowels /æ/ and /ʌ/ to the first language /a/ sound (Cebrian, 2019, 2021; Rallo Fabra & Romero, 2012). The perception of this vowel pair by Catalan–Spanish bilinguals with upper-intermediate to advanced English proficiency ranged from low to moderately high accuracy, depending on task difficulty (85.35% in an AXB discrimination task: Mora & Fullana, 2007; 31.0%–55.2% in a nonword identification task: Carlet & Cebrian, 2019; and 81.2%–86.7% in a real-word identification task: Carlet & Cebrian, 2014). As for production, even advanced Catalan–Spanish learners of English tended to use vowel duration rather than formant frequencies to distinguish the two vowels (Mora & Fullana, 2007). According to native American English speakers' perceptual judgments, Catalan–Spanish bilingual speakers tend to mispronounce low-front /æ/ as open-mid-front /ɛ/ (Rallo Fabra & Romero, 2012). For accurate production of /æ/, their formant values are expected to be higher for both F1 and F2. As for the English mid-back /ʌ/, its formant values are close to the Catalan and Spanish native open-central /a/, which results in the inaccurate production of /ʌ/ by Catalan–Spanish bilinguals (Rallo

Figure 1 From left to right, the lip hand gestures for /æ/ and /ʌ/, and the tongue hand gestures for /æ/ and /ʌ/.

Fabra & Romero, 2012). Thus, after training, we expect to observe a more accurate production of /ʌ/ with a lower F1 and F2.

In order to help Catalan–Spanish bilinguals to learn /æ/ and /ʌ/, three training conditions were designed, with the aim of providing the important articulatory information for the target vowels:

- **no-gesture (NG) condition**: The accessible articulatory information was the visible lip aperture.
- **lip hand gesture (LG) condition**: This provided training with hand gestures mimicking the lip aperture, where the accessible articulatory information was the lip aperture itself and the lip hand gesture.
- **tongue hand gesture (TG) condition**: This provided training with gestures mimicking the tongue backness, where the accessible articulatory information was the lip aperture and the tongue hand gesture.

Figure 1 shows the lip and tongue hand gestures used in our study. They were adapted from those described by Rusiewicz and Rivera (2017) and Chan (2018), respectively. Specifically, for the lip hand gesture, the thumb and fingers of one hand represented the lower and upper lips; thus, a large distance between the thumb and the four fingers mimicked the big lip aperture for /æ/, and a smaller distance the smaller lip aperture for /ʌ/. For the tongue hand gesture, following Chan (2018), two hands were used to represent the tongue position within the mouth: While one hand served as a fixed reference point, the other hand representing the tongue moved to show the relative tongue position. This allows for a more intuitive and easily understandable gesture form that minimizes spatial ambiguity. Specifically, the left hand was kept horizontal, palm downward, to represent the roof of the mouth as a fixed reference, while the right hand, held below the left and also palm downward, represented the respective target tongue positions during vowel production. The right hand was first held under the left hand to represent a neutral

tongue position in the mouth. Then, for the production of the front vowel /æ/, the right hand was moved forward by extending the arm further under the left arm, whereas for the production of the mid-back /ʌ/, the right hand was pulled backwards away from the neutral position. The decision to put both hand gestures in front of the body rather than next to the face was made to ensure comparability between the two gesture conditions as well as to maintain clear visibility of facial cues, enabling participants to benefit from both types of visual cue (i.e., facial articulatory cues and hand gestures) during training.

On the basis of previous literature, we hypothesized that both types of gestures should enhance learning compared to the NG condition, but hand gestures encoding visible articulatory features would be more beneficial for learning than hand gestures encoding nonvisible features. We thus expected to find the following hierarchy of training effects on both L2 perception and production:

> Training with hand gestures encoding visible articulatory features > training with hand gestures encoding nonvisible articulatory features > training without any gestures.

Method

Participants

Ninety-nine Catalan–Spanish bilinguals (84 females, $M_{Age} = 19.7$ years, $SD = 1.8$) who were 1st- and 2nd-year undergraduate students doing a degree in English language and literature at a public university in Barcelona volunteered to participate in this study. None of them reported having any hearing or speech impairments. We randomly assigned the participants to one of the following three conditions, described above: (a) the NG condition ($n = 33$, 26 females), (b) the LG condition ($n = 33$, 28 females), and (c) the TG condition ($n = 33$, 30 females).

Participants completed a questionnaire that provided us with background information about their age, their age of acquisition of English, the number of hours per week of instruction in English inside and outside of university, and the number of weeks they had spent studying English abroad. To examine whether there were any significant differences among the three training groups in terms of these variables, we performed the nonparametric Kruskal–Wallis test for each of the individual measures. The results showed no significant differences among the three groups in any of the individual measures (all ps $> .05$; see Table 1).

Table 1 Demographic characteristics of participants in the three conditions, with results of nonparametric analysis of variance tests

Characteristic	NG *M* (*SD*)	LG *M* (*SD*)	TG *M* (*SD*)	χ^2	*p*
Age	19.55 (1.6)	19.85 (1.9)	19.67 (2.0)	0.53	.769
Age of acquisition of English	5.18 (2.2)	5.12 (1.82)	5.21 (1.9)	0.22	.896
Hours per week of instruction in English in university	4.75 (10.4)	3.99 (5.0)	3.35 (3.1)	0.25	.884
Hours per week of instruction in English outside of university	1.06 (1.8)	1.13 (2.0)	1.66 (2.0)	2.22	.330
Weeks spent studying English abroad	2.79 (5.5)	7.03 (27.7)	4.67 (17.4)	0.17	.919

Note. NG = no-gesture condition; LG = lip hand gesture condition; TG = tongue hand gesture condition.

Materials

A male native speaker of American English[1] was video-recorded producing the materials for the training session and audio-recorded producing the testing materials. All the audiovisual materials for the training session were recorded in a broadcasting studio using a PDM660 Marantz professional portable digital recorder and a Rode NTG2 condenser microphone. Video clips were later edited in Adobe Premiere Pro CS6 and uploaded to Tobii Pro Lab, a platform used to conduct eye-tracking experiments that allowed us to create timelines with video stimuli and text instructions. All the audio materials for the testing session were recorded in a soundproof room using a Zoom H4n Pro recorder and a Shure SM35 headset condenser microphone. Audio clips were later edited using Audacity 2.4.2.

Audiovisual Stimuli for the Training Session

In order for participants to improve their perception and production of the target vowel pair /æ/ and /ʌ/, we deemed it crucial for them to have a clear understanding of the articulatory features involved. To this end, the instructor was filmed describing the differences between the two in terms of lip aperture and tongue position while at the same time explaining and demonstrating the respective hand gestures he would use in the two gesture conditions. This video material was then edited to produce three different versions of an introductory segment for the training video. All three versions included the instructor explaining the vowel contrast, but the version intended for the NG group did not

include demonstrations of the hand gestures, whereas the versions intended for the hand gesture groups included segments demonstrating only the hand gesture with which they would be trained.

To prepare the video materials to be used for training, we selected six consonant–vowel–consonant (CVC) minimal pairs in English that contrasted only in the vowels /æ/–/ʌ/ (e.g., *bag* vs. *bug*) and embedded them in 12 short sentences (see Appendix S1 in the Supporting Information online for the complete set of words and sentences). The instructor was video-recorded saying each of the words separately and then each sentence with the words embedded, with the shot framed around his face and upper torso so that his mouth and hand movements would be clearly visible to the viewer. For the NG condition, the instructor stood still and produced the training stimuli without any hand movements. For the LG and TG conditions, the instructor produced the lip or tongue hand gestures while producing the target sounds embedded in the words and sentences. Before the recording, the instructor was briefly trained to perform the gestures properly and practiced until he could perform them proficiently while pronouncing the related vowel. A total of 72 video clips were obtained in this fashion (12 words × 3 conditions + 12 sentences × 3 conditions). To avoid the potential influence of hand gestures on speech production, we superposed the audio tracks from the video clips in the NG condition over the audio tracks of the LG and TG videos. This also ensured that there would be no variation in the audio input that participants heard across conditions.

The various video clips were uploaded to Tobii Pro Lab software to create the final training videos for each condition. In each condition, the explanatory introduction preceded the training segment. Each word was trained as follows. First, a black screen appeared with the word in white text in the center of the screen. This was followed by the clip of the instructor saying the word accompanied by gesture or not, depending on the experimental condition. A black screen then appeared with the carrier sentence in white text, and this in turn was followed by a clip of the instructor saying the sentence, with or without gestures. This procedure was followed for each of the 12 word + sentence training items. Each full 12-item sequence was repeated in the training video a total of three times, with items appearing in a different order each time. Thus, participants were exposed to the training materials three times. In total, the duration of each full training video was about 15 min.

Stimuli for the Testing Sessions

Participants' perception and production of the target sounds were tested before, immediately after, and 1 week after the training session (i.e., in a

pretest, posttest, and delayed posttest). The three testing sessions consisted of the same four tasks in the following order: a paragraph-reading task, a picture-naming task, and a word-imitation task for speech production; and a word-identification task for speech perception (test materials are reproduced in Appendix S2 in the Supporting Information online). We presented the production tasks to the participants using an automatically timed PowerPoint, the last slide of which contained a link to the perception task on the Alchemer platform (https://www.alchemer.com).

For the paragraph-reading task, a short paragraph was adapted from the English phonetic textbook *Phonetic Words and Stories Book 5* (https://www.soundcityreading.net). The paragraph included a set of 14 words containing the target vowels /æ/ or /ʌ/ (e.g., *magic*, *bunny*): seven words for each sound. Crucially, none of these words had appeared in the training video.

For the picture-naming task, 10 common English words with everyday meanings were selected (e.g., *apple*, *bus*). Five of them contained /æ/, and the other five contained /ʌ/. We ensured that the chosen words were familiar to our student participants based on their proficiency level, as confirmed by their English teachers. None of the words had appeared in the training materials. Black and white pictures were then found that matched each word. Each picture was displayed on one slide, which was visible to participants for 5 s, giving them sufficient time to name what they saw.

For the word-imitation task, six English CVC minimal pairs contrasting only in /æ/–/ʌ/ were selected (e.g., *cab–cub*). Half of them had been included in the training materials, and the other half had not. The same instructor audio-recorded these 12 words, and each of the recordings was embedded in one presentation slide. In each trial, the PowerPoint automatically played the audio file, after which the word "Repeat" appeared on the screen for 3 s, during which interval the participant was expected to repeat the word.

For the word-identification task (to test perception), a set of six English CVC minimal pairs with the /æ/–/ʌ/ contrast were selected (e.g., *cat–cut*), none of which had appeared in the word-imitation task but half of which had been included in the training session. To increase the difficulty of the task, we added three CVC words containing /æ/ and three CVC words containing /ʌ/ that were not minimal pairs. This gave us a total of 18 words. The same instructor was recorded saying them, and then the resulting audio files were uploaded to Alchemer to build the survey. We created 18 internal pages, and the software randomized the order. Each page presented one audio recording of a word (e.g., *cat*) along with two minimal-pair words in written form that differed only in /æ/ versus /ʌ/ (e.g., *cat* vs. *cut*). Participants were tasked with

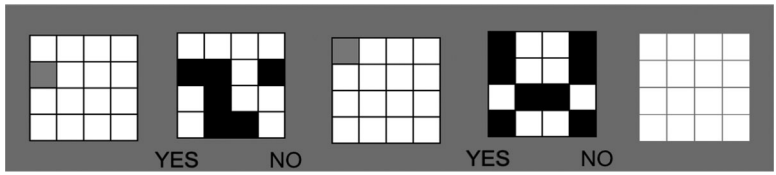

Figure 2 Example of a two-sequence trial from a visuospatial working memory test. Test takers are shown each image separately, in sequence. The squares shown as gray in the first and third images from the left here are shown in red in the test.

listening to the audio and subsequently choosing the correct option. As noted earlier, the link to this identification survey was embedded in the final slide of the PowerPoint.

It is worth noting that orthography may influence L2 speech sound learning (see Hayes-Harb & Barrios, 2021, for a review). In our study, considering the more complex grapheme-to-sound mappings in English compared to Catalan and Spanish, we carefully controlled the orthography of the words used in our training and testing stimuli. In particular, we consistently used the letter "a" for the /æ/ sound and the letter "u" for the /ʌ/ sound. In this way we aimed to mitigate any potential orthographic effects on the learning outcomes at the item level.

Materials for the Control Measures

In order to control for any potential differences between groups, we gave the participants two additional tasks to perform immediately after they completed the posttest. These consisted of the questionnaire described above (in the Participants section), related to their contact with L2 English, and a test of visuospatial working memory (VSWM), used because this variable has been shown to be positively related to the effects of gestural training during math instruction (Aldugom et al., 2020). The VSWM test took the form of an adapted symmetry span task (Blacker et al., 2017), hosted on the Pavlovia platform (https://pavlovia.org) and consisting of from two to six sequences of spatial recall prompts with interleaved symmetry judgment tasks. Figure 2 illustrates a two-sequence trial. In such a two-sequence trial, test takers were first shown a spatial recall prompt consisting of a 4 × 4 matrix with one red square. They were then shown a 4 × 4 matrix of black and white squares and asked to indicate if the design was symmetrical or not. This was followed by another spatial recall prompt showing one red square, in turn followed by another symmetry task. Finally, the test taker was shown a blank matrix and asked to indicate the

location of the two red squares they had seen previously and the order in which they had seen them. Each sequence was presented twice with varying prompts. As the task advanced, it became progressively more challenging, starting with two sequences, then proceeding to three, four, five, and finally six sequences. In a six-sequence trial, test takers saw six separate sequences of spatial recall prompt and symmetry test and were asked to indicate the location of all six red squares on the final matrix, in the correct order.

All the video clips used in the training session and slides used during the testing sessions, as well as the databases and statistical analyses reported below, are available via the Open Science Framework (https://osf.io/mnfxb/?view_only=f31915f8ffd04d9e93cd2ce0df961ab9).

Procedure

Before starting the experiment, participants signed a consent form to authorize the use of their data for academic purposes. All the training and testing tasks were conducted individually on a laptop in a soundproof room under the supervision of the first author of the study (the experimenter). Figure 3 provides an overview of the experimental procedure.

To begin the experiment, participants first completed the pretest (10 min), which consisted of paragraph-reading, picture-naming, word-imitation, and word-identification tasks, all delivered by means of the timed PowerPoint presentation described above. In the paragraph-reading task, they were instructed to read the text aloud. In the picture-naming task, they saw 10 pictures and had to say the word that each picture depicted within 5 s. In the word-imitation task, participants had to listen to each audio and orally repeat the testing word within 3 s of having seen the "Repeat" prompt. Throughout the three production tasks, the participants' speech output was recorded by a Zoom H4n Pro digital recorder using a Shure SM35 headset condenser microphone with the sample rate/resolution set at 44.1 kHz/16 bits. Recording was not necessary for the word-identification task, since for this task participants were linked on the last slide of the PowerPoint presentation to the online survey, where they heard a word and then identified by means of a cursor click the word they had heard.

Following the pretest, the experimenter randomly assigned participants to one of the three training conditions. The participants then watched the training video created for that group on Tobii Pro Lab software that was connected to a portable Tobii Pro X2-60 eye tracker. The training session took approximately 15 min. Before participants proceeded from the introductory video to the training video, the experimenter took the time to engage in direct communi-

Figure 3 Experimental procedure. NG = no-gesture condition; LG = lip hand gesture condition; TG = tongue hand gesture condition.

cation with the participants to confirm that they thoroughly understood the content and the training procedure. All participants were instructed to view the videos silently without imitating either the speech or the gestures of the instructor. Participants' eye-tracking behavior during the training session was recorded, but this will be discussed in a separate article. After watching the videos, participants completed the posttest (10 min), which consisted of exactly the same tasks as the pretest. Following this, they performed the VSWM task and answered the language background questionnaire (10 min for both combined). The participants were asked to return to the lab 1 week after the pretest–training–posttest session in order to perform the delayed posttest, which consisted of the same tasks they had performed in the pre- and posttests. Nine of the 99 participants failed to do so.

Data Coding

Production

To check participants' pronunciation of the target vowels, the recordings from the three production tasks were imported to Praat (Boersma & Weenink, 2022). As described above, the production stimuli contained 14 words in the paragraph-reading task, 10 words in the picture-naming task, and 12 words in the word-imitation task. Each of the 99 participants was to have been recorded producing these 36 words on three occasions (pretest, posttest, and delayed posttest), to yield a total of 10,692 recorded items. However, some items were not obtained because nine participants did not participate in the delayed posttest (324 items) and because (due to technical issues) some participant output was not recorded (206 items) or overlapped with extraneous noise (108 items). We therefore obtained a total of 10,054 recorded items. The first author manually annotated these recordings for the target sound /æ/ or /ʌ/ in Praat. During the annotation, 217 items containing mispronunciation (i.e., where the participants produced a nontarget word, such as *band* for *flag*, *peach* for *apple*, or *mug* for *cup*) were identified and thus excluded from the analysis. Finally, for the remaining 9,837 items, the first and second formants (F1 and F2) at the midpoints of the target vowels /æ/ and /ʌ/ were extracted from Praat using a script. The first author manually checked each item and corrected any abnormal formant values.

Perception

Participants' responses to the word-identification task were automatically assessed by the Alchemer application using a binary coding method: A correct answer scored 1, and an incorrect answer 0. The score for each item was then

exported from Alchemer and labeled as "Identification score." Since nine participants did not participate in the delayed posttest and one participant did not complete the perception task at the delayed posttest, a total of 5,166 identification scores were thus obtained.

Symmetry Span Task

Each participant's VSWM score was calculated by adding up the number of times that the participant managed to recall the correct position of a red square in the order in which it had appeared, a perfect score being 40 (Blacker et al., 2017).

Statistical Analyses

First, a parametric analysis of variance (ANOVA) was run to check whether there existed differences between groups in terms of VSWM. The results showed no statistically significant differences in this respect ($F(2,95) = 0.31$, $p = .737$, $\eta^2 = 0.006$).

To check for any effects of training on perception and production performance, we performed a binomial generalized linear mixed model for the identification score (binary variable) from the word-identification task and a set of linear mixed models (LMMs) for the F1 and F2 values (continuous variables) from the word-imitation task, the picture-naming task, and the paragraph-reading task using the lme4 package, Version 1.1-29 (Bates et al., 2015).

For our analysis of the identification score, the fixed effects were test (three levels: pretest, posttest, delayed posttest), condition (three levels: NG, LG, TG), and their interaction. As for the random effects, we built the model with the most complex random structures under the function buildmer() from the buildmer package, Version 2.8 (Voeten, 2022), and the random structures of the best fitting model included random intercepts for participant and item.

For our analysis of vowel quality, the six models (3 tasks × 2 formant values) involved the same fixed effects, which were test (three levels: pretest, posttest, delayed posttest), condition (three levels: NG, LG, TG), and vowel (two levels: /æ/, /ʌ/), as well as their interactions. As vowel quality values may vary depending on the speaker and the phonetic context, we included by-participant and by-item random intercepts in the random structures. The most complex model including random slopes for any repeated measures were built in buildmer(), which gave the output of the best random structures for the six models. The final structures that best fit our data were obtained and are reported below in the tables of the LMM results.

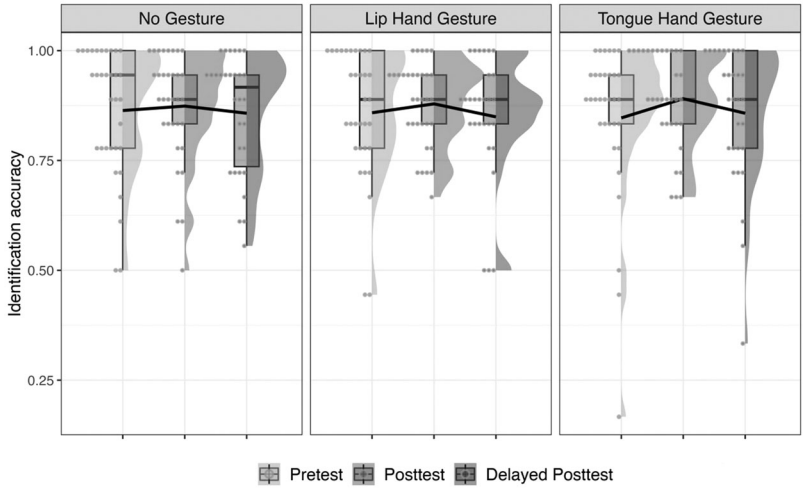

Figure 4 Boxplot and raincloud plot displaying participants' mean identification accuracy across condition and test in the word-identification task. The dots indicate the means of individual participants, and the black lines show the changes in group means over tests.

In all models, the significance values for the fixed effects with Type II Wald chi-square tests were calculated using the function Anova() from the car package, Version 3.0-13 (Fox & Weisberg, 2019). The post hoc pairwise comparisons were conducted with Bonferroni adjustment using the emmeans package, Version 1.7.4-1 (Lenth, 2022), which included Cohen's *d* as the effect size measure.

Results

Perception Performance

The mean word-identification accuracy for each of the three tests across conditions is plotted in Figure 4, and the results of the generalized linear mixed model are shown in Table 2. We obtained only a significant main effect of test. Post hoc results showed that for all three groups, perception accuracy improved immediately after training ($d = 0.28$, 95% CI [0.06, 0.49]; $p = .036$) but decreased again after 1 week ($d = -0.28$, 95% CI [−0.05, −0.50]; $p = .044$). As the results of the model showed no differences between groups, we can conclude that none of the conditions (including the two conditions involving hand gestures) was significantly more effective in helping learners improve their perception accuracy of /æ/ and /ʌ/.

Table 2 Results of the generalized linear mixed model for the word-identification task

	Identification score	
Fixed effects	χ^2	p
Test	7.84	.020
Condition	0.07	.963
Test × Condition	2.26	.689

Note. Model formula: Score ~ Test * Condition + (1 | Item) + (1 | Participant).

Production Performance

As English low-front /æ/ has higher F1 and F2 frequency values than the mid-back /ʌ/ but Catalan–Spanish bilingual learners of English produce smaller formant frequency differences compared to native speakers, we expected that after training, the F1 and F2 values of /æ/ produced by participants would increase whereas the two formant values of /ʌ/ would decrease. In the following, we will first report the results of the three LMMs for the F1 values in the word-imitation, picture-naming, and paragraph-reading tasks. Then we will report the results of the three LMMs for the F2 values in the three production tasks.

F1 Values: Lip Aperture

Figure 5 displays boxplots of the mean F1 values in hertz of /æ/ and /ʌ/ across conditions and tests for the three production tasks. Table 3 summarizes the results of the three LMMs.

For the word-imitation task, post hoc results of the significant Test × Vowel interaction showed that for all three groups in general, the F1 values of /æ/ remained unchanged (all ps $> .05$), but the F1 values of /ʌ/ decreased immediately after training ($d = -0.42$, 95% CI [−0.22, −0.63]; $p = .002$), and the decreased values were maintained after 1 week ($d = 0.32$, 95% CI [0.01, 0.63]; $p > .05$).

For the picture-naming task, post hoc results of the significant three-way interaction are as follows:

- **NG group**: The F1 values of /æ/ became higher immediately after training ($d = 0.47$, 95% CI [0.20, 0.74]; $p = .001$) and maintained the improvement 1 week later ($d = 0.14$, 95% CI [−0.23, 0.51]; $p > .05$), whereas the values of /ʌ/ showed no changes after training ($d = -0.09$,

Figure 5 Boxplot and raincloud plot displaying first formant (F1) values in hertz for the vowels /æ/ and /ʌ/ across condition and test in the word-imitation task (left panel), picture-naming task (center panel), and paragraph-reading task (right panel). The dots indicate the F1 means of individual participants, and the black lines show the changes in group means over tests. NG = no-gesture condition; LG = lip hand gesture condition; TG = tongue hand gesture condition.

Table 3 Results of the three linear mixed models for the first formant (F1) values in the word-imitation, picture-naming, and paragraph-reading tasks

	Word-imitation task		Picture-naming task		Paragraph-reading task	
Fixed effects	χ^2	p	χ^2	p	χ^2	p
Test	7.40	**.025**	15.31	**<.001**	7.05	**.029**
Vowel	59.93	**<.001**	16.34	**<.001**	3.43	.064
Condition	0.08	.959	5.44	.066	4.25	.119
Test × Vowel	14.58	**<.001**	60.55	**<.001**	20.61	**<.001**
Test × Condition	1.33	.856	7.32	.120	11.23	.024
Vowel × Condition	1.79	.408	1.02	.599	0.56	.757
Test × Vowel × Condition	3.74	.442	10.49	**.033**	14.50	**.006**

Note. Boldface indicates statistical significance at $\alpha = .05$. Model formulas: (a) word-imitation task: F1 ∼ Test ∗ Vowel ∗ Condition + (1 + Test ∗ Vowel | Participant) + (1 + Test | Item); (b) picture-naming task: F1 ∼ Test ∗ Vowel ∗ Condition + (1 + Test + Vowel | Participant) + (1 | Item); (c) paragraph-reading task: F1 ∼ Test ∗ Vowel ∗ Condition + (1 + Test + Vowel | Participant) + (1 + Test | Item).

95% CI [−0.36, 0.19]; $p > .05$) and became higher 1 week after the posttest ($d = 0.63$, 95% CI [0.25, 1.01]; $p = .003$).

- **LG group**: The F1 values of /æ/ became higher ($d = 0.33$, 95% CI [0.06, 0.60]; $p = .039$) and also maintained the improvement after 1 week ($d = 0.10$, 95% CI [−0.27, 0.48]; $p > .05$), and the F1 values of /ʌ/ decreased immediately after training ($d = -0.74$, 95% CI [−0.46, −1.02]; $p < .001$) but increased 1 week later ($d = 0.66$, 95% CI [0.28, 1.04]; $p = .001$).
- **TG group**: A significant change was observed only for the F1 values of /ʌ/, which increased 1 week after the posttest ($d = 0.62$, 95% CI [0.23, 1.00]; $p = .004$).

For the paragraph-reading task, post hoc results of the significant three-way interaction are as follows:

- **NG group**: No significant differences for either /æ/ or /ʌ/ across tests were observed (all ps $> .05$).
- **LG group**: Whereas F1 values of /æ/ showed no change (all ps $> .05$), the values of /ʌ/ decreased from pretest to posttest ($d = -0.57$, 95% CI

[−0.34, −0.81]; $p < .001$), although they increased 1 week later ($d = 0.75$, 95% CI [0.34, 1.15]; $p < .001$).

- **TG group**: Whereas no significant changes over time were observed for /æ/ (all ps $> .05$), the F1 values of /ʌ/ at delayed posttest were significantly higher than the pretest values ($d = 0.48$, 95% CI [0.08, 0.89]; $p = .045$).

Taken together, the three training conditions showed different effects with regard to adjusting the articulation of lip aperture for the target vowels /æ/ and /ʌ/, depending on the production task. First, in the word-imitation task, none of the conditions seems to have helped participants adjust their lip aperture for /æ/, but all three showed similar benefits for /ʌ/. Second, in the picture-naming task, whereas the TG condition did not help participants adjust their lip aperture for /ʌ/, the NG condition was helpful for /æ/, and the LG condition was helpful for both /æ/ and /ʌ/. Finally, in the paragraph-reading task, only the LG condition had positive effects on adjusting lip aperture for the /ʌ/ sound. These results suggest that the LG condition was more effective than the other two conditions in causing participants to adjust their lip aperture for the target vowels, especially in the picture-naming and paragraph-reading tasks.

F2 Values: Tongue Backness

Figure 6 displays boxplots of the mean F2 values in hertz for /æ/ and /ʌ/ across conditions and tests in the three production tasks. Table 4 summarizes the results of the three LMMs.

For the word-imitation task, post hoc results of the significant Test × Vowel interaction revealed that the three training conditions did not yield any significant change in the F2 values for either /æ/ or /ʌ/ over time (all ps $> .05$).

For the picture-naming task, post hoc results of the significant Test × Vowel interaction revealed that all three conditions helped to increase the F2 values of /æ/ after training ($d = 0.29$, 95% CI [0.15, 0.43]; $p < .001$) and to maintain this improvement 1 week later ($d = -0.03$, 95% CI [−0.17, 0.12]; $p > .05$). As for the /ʌ/ sound, the F2 values increased 1 week after the posttest ($d = 0.28$, 95% CI [0.12, 0.43]; $p < .001$) but showed no significant difference from the values at pretest ($d = 0.15$, 95% CI [−0.01, 0.30]; $p > .05$). In addition, the significant Test × Condition interaction revealed that regardless of the vowel, F2 values significantly decreased from the posttest to the delayed posttest in both the NG group ($d = -0.32$, 95% CI [−0.14, −0.49]; $p < .001$) and the TG group ($d = -0.23$, 95% CI [−0.05, −0.41]; $p = .021$). In contrast, the LG group did not show any significant change in F2 values (all ps $> .05$).

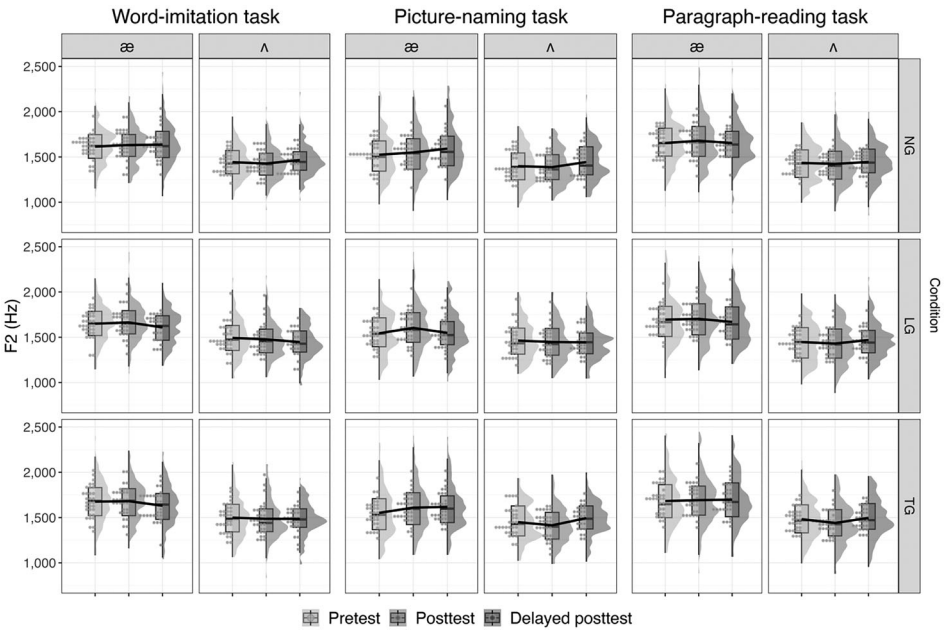

Figure 6 Boxplot and raincloud plot displaying second formant (F2) values in hertz for the vowels /æ/ and /ʌ/ across condition and test in the word-imitation task (left panel), picture-naming task (center panel), and paragraph-reading task (right panel). The dots indicate the F2 means of individual participants, and the black lines show the changes in group means over tests. NG = no-gesture condition; LG = lip hand gesture condition; TG = tongue hand gesture condition.

Table 4 Results of the three linear mixed models for the second formant (F2) values in the word-imitation, picture-naming, and paragraph-reading tasks

	Word-imitation task		Picture-naming task		Paragraph-reading task	
Fixed effects	χ^2	p	χ^2	p	χ^2	p
Test	1.68	.431	18.59	**<.001**	1.12	.572
Vowel	20.90	**<.001**	5.70	**.017**	13.84	**<.001**
Condition	1.15	.563	1.62	.444	2.12	.347
Test × Vowel	15.59	**<.001**	21.17	**<.001**	12.69	**.002**
Test × Condition	6.75	.150	20.89	**<.001**	0.94	.919
Vowel × Condition	0.49	.783	1.10	.577	0.19	.909
Test × Vowel × Condition	2.86	.581	3.42	.490	1.31	.860

Note. Model formulas: (a) word-imitation task: F2 ∼ Test ∗ Vowel ∗ Condition + (1 + Test + Vowel | Participant) + (1 | Item); (b) picture-naming task: F2 ∼ Test ∗ Vowel ∗ Condition + (1 + Vowel | Participant) + (1 | Item); (c) paragraph-reading task: F2 ∼ Test ∗ Vowel ∗ Condition + (1 | Item) + (1 + Vowel + Test | Participant).

For the paragraph-reading task, post hoc results of the significant Test × Vowel interaction revealed that the F2 values for the /æ/ remained unchanged across the three tests (all ps > .05). As for /ʌ/, F2 values did not change from pretest to posttest ($d = -0.10$, 95% CI [−0.23, 0.03], $p > .05$). Although F2 increased 1 week after training ($d = 0.20$, 95% CI [0.06, 0.35]; $p = .009$), the delayed posttest did not show significantly different F2 values from the pretest ($d = 0.10$, 95% CI [−0.04, 0.24]; $p > .05$).

Taken together, the results show that the three training conditions showed equally limited effects on adjusting tongue position for the target vowels /æ/ and /ʌ/. Whereas all three training conditions failed to help participants to retract their tongues for the mid-back vowel /ʌ/ in all three production tasks, they helped participants to attain the front tongue position for the low-front /æ/ in the picture-naming task.

Discussion

The present study examined the benefits of observing hand gestures mimicking articulatory features of L2 English /æ/–/ʌ/ for the learning of the target vowels by Catalan–Spanish learners of English. The main goal of the study was to determine the importance of the visibility of the target articulation during audiovisual phonetic training with hand gestures. Specifically, the two types of

hand gestures used in this study encoded articulatory features that differed in their visual accessibility: The lip hand gesture encoded the visible lip aperture, and the tongue hand gesture mimicked the nonvisible tongue position. We hypothesized a hierarchy of training effects on L2 perception and production, as follows:

> Training with hand gestures encoding visible articulatory features > training with hand gestures encoding nonvisible articulatory features > training without any gestures.

Production of L2 Vowel Contrasts

Our production results revealed that the LG condition outperformed both the NG condition and the TG condition in helping learners to adjust the lip aperture of the mid-back vowel /ʌ/ immediately after training in both the paragraph-reading and picture-naming tasks. They also helped with the lip aperture of the low-front vowel /æ/ in the paragraph-reading task immediately after training as well as after 1 week. By contrast, the TG condition did not yield more benefits in adjusting the lip aperture or tongue backness than the NG condition.

These results confirm and expand some previous results reported by Hoetjes and van Maastricht (2020), which also showed the positive effects of using gestures mimicking lip shape as opposed to gestures representing the tongue. However, the null effect of the tongue hand gesture is inconsistent with some studies that showed the effectiveness of using gestures mimicking tongue shape in both classroom (Lan & Wu, 2013) and clinical (Rusiewicz & Rivera, 2017) settings. These two studies not only asked participants to perform hand gestures, but also incorporated immediate feedback from either teachers or clinicians. By contrast, participants in both Hoetjes and van Maastricht's study and the current study watched the training video but received no feedback on their performance. For attaining new articulation patterns that are not readily visible like tongue position, feedback might be key for successful speech motor learning to occur (e.g., see Levitt & Katz, 2007, for training involving the tongue tip using augmented visual feedback).

Importantly, our results related to visibility also raise the question of why it is that hand gestures mimicking lip and tongue features show different effects at the production level. According to the dual coding theory (Clark & Paivio, 1991), learners are in principle expected to benefit from the processing of visual information. But our results show that not all types of visual input coming from hand gestures are useful. During multimodal phonetic training, apart from accessing information from hand gestures, learners are exposed to

acoustic information and also to visual information about the mouth and face. Thus, learners are exposed to two visual inputs at the same time. In the LG condition, the lip hand gesture is perceptually congruent with the visible shape of the lip. Thus, the two visual inputs complement each other by referring to the same articulatory information, which can raise the learners' articulatory awareness and enhance the stability of their phonological–motor mapping. However, in the TG condition, the tongue hand gesture is not matching any visually accessible articulatory information. In fact, the two visual inputs available to the learner (i.e., the hands and the lip movements) contain contrasting information, as humans "are normally unaccustomed to seeing the movements of the intra-oral articulators" (Wik & Engwall, 2008), and there exists a predominance of reading the lips over the tongue (Badin et al., 2010). Thus, the TG condition may increase the demands on learners of processing speech articulation and thus not help in phonological learning. All in all, including visual inputs of hand gestures encoding articulatory information in phonetic training might be more beneficial than not including them, provided that the hand gestures can be easily matched with the target articulation.

Interestingly, participants' improvement in their vowel production during the paragraph-reading task and the picture-naming task contrasted with their performance in the word-imitation task, in which no group differences were found in terms of changing either F1 or F2 values. In the word-imitation task, all training conditions helped participants to adjust their lip aperture for /ʌ/ and maintain their learning performance for 1 week, whereas no beneficial effects were observed for the tongue positions required for /ʌ/ or for either of the two articulatory features with regard to /æ/. The limited effects of hand gestures on the imitation of nonnative sounds are inconsistent with previous findings showing that gestures mimicking durational or aspiration features improved learners' imitation of L2 sounds (Li et al., 2020; Xi et al., 2020). Whereas the participants in the previous two studies were naive learners with no experience in the target L2 (Li et al., 2020; Xi et al., 2020), our participants had an intermediate-level mastery of the L2. Therefore, the imitation task in our study may have been too easy to enable detection of pronunciation improvement after training. This observation is consistent with the findings of Ozakin et al. (2023), who also noted that more challenging tasks, such as paragraph reading, were more sensitive in detecting the different training effects across groups. Importantly, for advanced learners, researchers have suggested that speech production in controlled tasks, such as imitation, may not accurately reflect learners' ability in real-world communication (Lee et al., 2015) and does not necessarily reflect their productive knowledge of nonnative sounds (Llompart & Reinisch, 2019).

In addition, an important effect of visibility was found in the NG condition. Although participants were informed of the importance of both lip aperture and tongue backness for accurately producing the target vowels, the access to the visible articulatory information conveyed by the lips in the NG condition helped learners to adjust their lip aperture appropriately for the target vowels immediately after training (benefits were observed for /æ/ in the picture-naming task, $d = 0.47$, and /ʌ/ in the word-imitation task, $d = 0.42$). By contrast, since audiovisual phonetic training could not allow learners to access information about internal tongue position, the training effects on the tongue backness feature were relatively smaller ($d = 0.29$ for /æ/ and $d = 0.28$ for /ʌ/ in the picture-naming task). This result confirms and expands the long-established line of work demonstrating that having visual access to articulatory features positively impacts phonetic learning (Hardison, 2003; Hazan et al., 2005; Inceoglu, 2016).

Perception of L2 Vowel Contrasts

For the perception of L2 sounds, the results of the word-identification task showed that participants in all three training conditions improved their perceptual accuracy for English /æ/ and /ʌ/, but the improvement was not maintained after 1 week. Thus, neither of the two gesture conditions yielded more perceptual identification improvement than training without gestures. These results are in line with previous studies showing that the integration of hand gestures into audiovisual phonetic training yielded limited effects on L2 perception in terms of encoding acoustic features (Hirata & Kelly, 2010; Hirata et al., 2014; Li et al., 2020, 2021; Xi et al., 2020). One possible explanation of the null effects is that in some cases either explicit information about articulation (Linebaugh & Roche, 2015) or the observation of facial cues (Hazan et al., 2005) is sufficient to establish new categories in the L2 learners' phonological system. The articulatory information encoded by hand gestures may be redundant for L2 perception learning. Moreover, the accuracy of perception of English /æ/ and /ʌ/ by the participants in the present study was already quite high at pretest (NG: 86.36%; LG: 85.86%; TG: 85.68%), which left little room for improvement. It might well be that using an identification task involving pseudowords, as done by Carlet and Cebrian (2019), or using real words embedded in carrier sentences, as done by Li et al. (2020), would have helped to increase the difficulty of the task and thus made the results of training more salient.

Implications for L2 Pronunciation Instruction

The current study has several implications for the importance of gesture use in L2 pronunciation instruction. First, the LG condition allowed us to observe an immediate training effect with a small effect size for the pronunciation of /æ/ ($d = 0.33$ in the picture-naming task and $d = 0.06$ in the paragraph-reading task), and a medium effect size for the pronunciation of /ʌ/ ($d = 0.74$ in the picture-naming task and $d = 0.57$ in the paragraph-reading task). Interestingly, this gain seems to be larger than the gains obtained from nonmultimodal training methods, with the mean Cohen's *d* of the immediate training effect being 0.54 (see Sakai & Moorman, 2018, for a meta-analysis of perception-based phonetic training). This comparison provides evidence supporting the positive role of hand gestures in boosting L2 pronunciation training. Second, even though the training involved listening to minimal pairs of words, a clear generalization effect was found in the pronunciation tests at the posttest in the LG condition. Specifically, the positive effects were found in the paragraph-reading task and spontaneous picture-naming task, both of which contained items that were not included in the training. Thus, the improvement observed after a brief audiovisual phonetic training session involving hand gestures was not merely attributable to memorization, since robust phonological learning did occur.

Limitations and Future Research

Several limitations of this study and further research questions can be identified. First, although a short training session with hand gestures encoding visible lip aperture helped learners to improve their pronunciation of the target vowels, they were not always able to maintain their productive learning outcomes over time. In our view, future research with a longitudinal design could usefully be conducted to test the retention of phonological learning effects over multiple training sessions. Second, many individual variables can potentially influence the effectiveness of phonetic training involving hand gestures for L2 sound acquisition. For example, an individual's sensitivity to visual cues (Sennema et al., 2003) and attention to hand gestures may be related to their ability to process multimodal information (Kandana Arachchige et al., 2021). Future studies could specifically consider how individual variables interact with the effects of gesture-based phonetic training. Third, as pointed out by an anonymous reviewer, the degree of lip aperture, although strongly correlated with F1 values, is not the only variable influencing F1 values. Other variables, such as the position of the tongue root or the degree of lip roundness, could also impact F1. However, our training directed the learners' attention to lip aperture (through instruction in the introductory video

or through the lip hand gestures in the LG group), and this makes it more plausible that the observed F1 changes after training resulted from adjustments in lip aperture. It remains possible that certain individual participants might have adjusted their F1 values through modifications in tongue height or in lip roundedness, rather than by manipulating lip aperture, but it is improbable that the entire group did so, as they were not instructed to. Undoubtedly, future studies could enhance the methodology by incorporating articulatory phonetic tools like lip cameras, which would permit a direct assessment of lip aperture.

Conclusion

The present study has shown that audiovisual phonetic training with hand gestures encoding articulatory features is more effective in helping learners to improve their L2 vowel pronunciation than training without hand gestures, but only when the hand gestures reinforce articulatory features that are already visible. It highlights the importance of visual accessibility of articulatory features encoded by hand gestures in the learning of L2 speech sounds and also adds further evidence in favor of the embodied cognition paradigm for language learning (for a review, see Jusslin et al., 2022). These findings have clear applications in educational settings. Hand gestures are pedagogical tools that can be easily integrated into L2 classroom sessions, but the choice of gestures and what they are intended to represent will influence their ability to optimally boost phonological language learning.

Final revised version accepted 10 January 2024

Open Research Badges

This article has earned Open Data and Open Materials badges for making publicly available the digitally-shareable data and the components of the research methods needed to reproduce the reported procedure and results. All data and materials that the authors have used and have the right to share are available at https://osf.io/mnfxb/?view_only=f31915f8ffd04d9e93cd2ce0df961ab9. All proprietary materials have been precisely identified in the manuscript.

Note

1 As pointed out by an anonymous reviewer, Spanish schools usually follow British English in their curricula of English courses. Despite this, a native American English speaker was selected as the instructor in this study. This is because in our specific university population, students were exposed to a variety of English accents in class, namely Catalan/Spanish-accented L2 English or native English. In the

academic year in which the experiment took place, there were eight English teachers, of whom three were native American English speakers, one was an Irish English speaker, and the remaining four were Catalan/Spanish speakers. In other words, participants in our study were acquainted with the American English accent.

References

Aldugom, M., Fenn, K., & Cook, S. W. (2020). Gesture during math instruction specifically benefits learners with high visuospatial working memory capacity. *Cognitive Research: Principles and Implications*, *5*, 27. https://doi.org/10.1186/s41235-020-00215-8

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. https://doi.org/10.1515/rela-2016-0020

Badin, P., Tarabalka, Y., Elisei, F., & Bailly, G. (2010). Can you "read" tongue movements? Evaluation of the contribution of tongue display to speech understanding. *Speech Communication*, *52*(6), 493–503. https://doi.org/10.1016/j.specom.2010.03.002

Barriuso, T. A., & Hayes-Harb, R. (2018). High variability phonetic training as a bridge from research to practice. *The CATESOL Journal*, *30*(1), 177–194. http://www.catesoljournal.org/wp-content/uploads/2018/03/CJ30.1_barriuso.pdf

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

Blacker, K. J., Weisberg, S. M., Newcombe, N. S., & Courtney, S. M. (2017). Keeping track of where we are: Spatial working memory in navigation. *Visual Cognition*, *25*(7–8), 691–702. https://doi.org/10.1080/13506285.2017.1322652

Blackwood Ximenes, A., Shaw, J. A., & Carignan, C. (2017). A comparison of acoustic and articulatory methods for analyzing vowel differences across dialects: Data from American and Australian English. *The Journal of the Acoustical Society of America*, *142*(1), 363–377. https://doi.org/10.1121/1.4991346

Boersma, P., & Weenink, D. (2022). *Praat: Doing phonetics by computer* (Version 6.1.51) [Computer software]. http://www.praat.org/

Carlet, A., & Cebrian, J. (2014). Training Catalan speakers to identify L2 consonants and vowels: A short-term high variability training study. *Concordia Working Papers in Applied Linguistics*, *5*, 85–98. https://www.researchgate.net/publication/318129279

Carlet, A., & Cebrian, J. (2019). Assessing the effect of perceptual training on L2 vowel identification, generalization and long-term effects. In A. M. Nyvad, M. Hejná, A. Højen, A. B. Jespersen, & M. H. Sørensen (Eds.), *A sound approach to*

language matters: In honor of Ocke-Schwen Bohn (pp. 91–119). Aarhus University. https://doi.org/10.7146/aul.322.218

Cebrian, J. (2019). Perceptual assimilation of British English vowels to Spanish monophthongs and diphthongs. *The Journal of the Acoustical Society of America*, *145*(1), EL52–EL58. https://doi.org/10.1121/1.5087645

Cebrian, J. (2021). Perception of English and Catalan vowels by English and Catalan listeners: A study of reciprocal cross-linguistic similarity. *The Journal of the Acoustical Society of America*, *149*(4), 2671–2685. https://doi.org/10.1121/10.0004257

Chan, M. J. (2018). Embodied pronunciation learning: Research and practice. *The Catesol Journal*, *30*(1), 47–68. http://www.catesoljournal.org/wp-content/uploads/2018/03/CJ30.1_chan.pdf

Clark, J. M., & Paivio, A. (1991). Dual coding theory and education. *Educational Psychology Review*, *3*(3), 149–210. https://doi.org/10.1007/BF01320076

Derwing, T. M., & Munro, M. J. (2015). *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research*. John Benjamins Publishing Company.

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3rd ed.). SAGE.

Fromkin, V. (1964). Lip positions in American English vowels. *Language and Speech*, *7*(4), 215–225. https://doi.org/10.1177/002383096400700402

Grauwinkel, K., Dewitt, B., & Fagel, S. (2007). Visual information and redundancy conveyed by internal articulator dynamics in synthetic audiovisual speech. *Proceedings of Interspeech 2007*, 706–709. https://doi.org/10.21437/Interspeech.2007-295

Hardison, D. M. (2003). Acquisition of second-language speech: Effects of visual cues, context, and talker variability. *Applied Psycholinguistics*, *24*(4), 495–522. https://doi.org/10.1017/S0142716403000250

Hayes-Harb, R., & Barrios, S. (2021). The influence of orthography in second language phonological acquisition. *Language Teaching*, *54*(3), 297–326. https://doi.org/10.1017/S0261444820000658

Hazan, V., Sennema, A., Iba, M., & Faulkner, A. (2005). Effect of audiovisual perceptual training on the perception and production of consonants by Japanese learners of English. *Speech Communication*, *47*(3), 360–378. https://doi.org/10.1016/j.specom.2005.04.007

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. https://doi.org/10.1044/1092-4388(2009/08-0243)

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. https://doi.org/10.1044/2014_JSLHR-S-14-0049

Hoetjes, M., & van Maastricht, L. (2020). Using gesture to facilitate L2 phoneme acquisition: The importance of gesture and phoneme complexity. *Frontiers in Psychology*, *11*, Article 575032. https://doi.org/10.3389/fpsyg.2020.575032

Hudson, N. (2011). *Teacher gesture in a post-secondary English as a second language classroom: A sociocultural approach* [Doctoral dissertation, University of Nevada Las Vegas]. UNLV Theses, Dissertations, Professional Papers, and Capstones. https://doi.org/10.34917/2432927

Inceoglu, S. (2016). Effects of perceptual training on second language vowel perception and production. *Applied Psycholinguistics*, *37*(5), 1175–1199. https://doi.org/10.1017/S0142716415000533

Jusslin, S., Korpinen, K., Lilja, N., Martin, R., Lehtinen-Schnabel, J., & Anttila, E. (2022). Embodied learning and teaching approaches in language education: A mixed studies review. *Educational Research Review*, *37*, Article 100480. https://doi.org/10.1016/j.edurev.2022.100480

Kandana Arachchige, K. G., Blekic, W., Simoes Loureiro, I., & Lefebvre, L. (2021). Covert attention to gestures is sufficient for information uptake. *Frontiers in Psychology*, *12*, Article 776867. https://doi.org/10.3389/fpsyg.2021.776867

Lan, Y., & Wu, M. (2013). Application of form-focused instruction in English pronunciation: Examples from Mandarin learners. *Creative Education*, *4*(9B), 29–34. https://doi.org/10.4236/ce.2013.49b007

Lee, J., Jang, J., & Plonsky, L. (2015). The effectiveness of second language pronunciation instruction: A meta-analysis. *Applied Linguistics*, *36*(3), 345–366. https://doi.org/10.1093/applin/amu040

Lenth, R. V. (2022). *emmeans: Estimated marginal means, aka least-squares means* (R package; Version 1.7.4-1) [Computer software]. https://CRAN.R-project.org/package=emmeans

Levitt, J. S., & Katz, W. F. (2007). Augmented visual feedback in second language learning: Training Japanese post-alveolar flaps to American English speakers. *Proceedings of Meetings on Acoustics*, *2*(1), Article 060002. https://doi.org/10.1121/1.2992054

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

Linebaugh, G., & Roche, T. (2015). Evidence that L2 production training can enhance perception. *Journal of Academic Language and Learning*, *9*(1), A1–A17. https://journal.aall.org.au/index.php/jall/article/view/326

Llompart, M., & Reinisch, E. (2019). Imitation in a second language relies on phonological categories but does not reflect the productive usage of difficult sound

contrasts. *Language and Speech*, *62*(3), 594–622. https://doi.org/10.1177/0023830918803978

Mora, J. C., & Fullana, N. (2007). Production and perception of English /i:/-/ɪ/and /æ/-/ʌ/ in a formal setting: Investigating the effects of experience and starting age. In J. Trouvain & W. J. Barry (Eds.), *Proceedings of the 16th International Congress of Phonetic Sciences* (pp. 1613–1616). http://www.icphs2007.de/conference/Papers/1594/1594.pdf

Muñoz, C. (2008). Symmetries and asymmetries of age effects in naturalistic and instructed L2 learning. *Applied Linguistics*, *29*(4), 578–596. https://doi.org/10.1093/applin/amm056

Ozakin, A. S., 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

Paivio, A. (1991). Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *45*(3), 255–287. https://doi.org/10.1037/h0084295

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, *24*(2), 175–184. https://doi.org/10.1121/1.1906875

Pillot-Loiseau, C., Kocjančič Antolík, T., & Kamiyama, T. (2013). Contribution of ultrasound visualisation to improving the production of the French /y/-/u/contrast by four Japanese learners. *Proceedings of the PPLC13*, 86–89. http://hal.archives-ouvertes.fr/hal-00862367

Rallo Fabra, L., & Romero, J. (2012). Native Catalan learners' perception and production of English vowels. *Journal of Phonetics*, *40*(3), 491–508. https://doi.org/10.1016/j.wocn.2012.01.001

Reyes, Y. P., & Hazan, V. (2021). English vowel perception by non-native speakers: Impact of audio and visual training modalities. *Onomazein*, *51*, 111–136. https://doi.org/10.7764/onomazein.51.04

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience*, *27*, 169–192. https://doi.org/10.1146/annurev.neuro.27.070203.144230

Rusiewicz, H. L., & Rivera, J. L. (2017). The effect of hand gesture cues within the treatment of /r/for a college-aged adult with persisting childhood apraxia of speech. *American Journal of Speech-Language Pathology*, *26*(4), 1236–1243. https://doi.org/10.1044/2017_AJSLP-15-0172

Sakai, M., & Moorman, C. (2018). Can perception training improve the production of second language phonemes? A meta-analytic review of 25 years of perception training research. *Applied Psycholinguistics*, *39*(1), 187–224. https://doi.org/10.1017/S0142716417000418

Sennema, A., Hazan, V., & Faulkner, A. (2003). The role of visual cues in L2 consonant perception. *15th International Congress of Phonetic Sciences*, 135–138. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2003/p15_0135.html

Smotrova, T. (2017). Making pronunciation visible: Gesture in teaching pronunciation. *TESOL Quarterly*, *51*(1), 59–89. https://doi.org/10.1002/tesq.276

Sullivan, J. V. (2018). Learning and embodied cognition: A review and proposal. *Psychology Learning and Teaching*, *17*(2), 128–143. https://doi.org/10.1177/1475725717752550

Voeten, C. C. (2022). *buildmer: Stepwise elimination and term reordering for mixed-effects* (R package; Version 2.4) [Computer software]. https://cran.r-project.org/package=buildmer

Wang, X., Hueber, T., & Badin, P. (2014). On the use of an articulatory talking head for second language pronunciation training: The case of Chinese learners of French. *Proceedings of the 10th International Seminar on Speech Production – ISSP 2014*, 449–452. https://hal.ird.fr/INPG/hal-00974342v1

Wik, P., & Engwall, O. (2008). Looking at tongues: Can it help in speech perception? *Proceedings of FONETIK 2008*, 57–60. https://www.academia.edu/28071988/Looking_at_tongues_can_it_help_in_speech_perception?uc-sb-sw=107199741

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review*, *9*(4), 625–636. http://view.ncbi.nlm.nih.gov/pubmed/12613670

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

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Accessible Summary

Appendix S1. Words and Sentences Used in the Training Session.

Appendix S2. Materials for the Testing Sessions.