

Research Article

OBSERVING AND PRODUCING DURATIONAL HAND GESTURES FACILITATES THE PRONUNCIATION OF NOVEL VOWEL-LENGTH CONTRASTS

Peng Li  *

Universitat Pompeu Fabra

Florence Baills

Universitat Pompeu Fabra

Pilar Prieto

Institució Catalana de Recerca i Estudis Avançats (ICREA)

Universitat Pompeu Fabra

Abstract

While empirical studies have shown the beneficial role of observing and producing hand gestures mimicking pitch features in the learning of L2 tonal or intonational contrasts, mixed results have been obtained for the use of gestures encoding durational contrasts at the perceptual level. This study investigates the potential benefits of horizontal hand-sweep gestures encoding durational features for boosting the perception and production of nonnative vowel-length contrasts. In a between-subjects experiment with a pretest–posttest design, 50 Catalan participants without any knowledge of Japanese practiced perceiving and producing minimal pairs of Japanese disyllabic words featuring vowel-length contrasts in one of two conditions, namely with gestures or without

The authors sincerely thank Misa Fukukawa, Yosuke Nakano, and Ingrid Vilà-Giménez for their participation in creating the materials and Yuan Zhang for her assistance in conducting the experiment. Many thanks also to Lorraine Baqué, Joan Carles Mora, and Kazuya Saito for their comments and suggestions on the first draft of this article. Special thanks also go to the two anonymous reviewers and the editors, Susan Gass and Elizabeth Huntley, for invaluable comments and feedback.

This research was supported by funding from the Spanish Ministry of Economy and Competitiveness (FFI2015-66533-P) and the Generalitat de Catalunya projects (2014 SGR-925 and 2017 SGR-971). The second author has a predoctoral research grant awarded by the Department of Translation and Language Sciences, Universitat Pompeu Fabra.

*Correspondence concerning this article should be addressed to Peng Li, Department of Translation and Language Sciences, Universitat Pompeu Fabra, Roc Boronat 138, 08018 Barcelona, Spain. E-mail: peng.li@upf.edu.

them. Pretest and posttest consisted of the completion of identical vowel-length identification and imitation tasks. The results showed that while participants improved equally at posttest across the two conditions in the identification task, the Gesture group obtained a larger improvement than the No Gesture group in the imitation task. These results corroborate the claim that producing hand gestures encoding prosodic properties of speech may help naïve learners to learn novel phonological contrasts in a foreign language.

INTRODUCTION

In the last few decades, a growing body of research has shown that hand gestures play an important supporting role not only in the context of first language (L1) learning (e.g., Goldin-Meadow, 2010, 2011) but also in that of second language (L2) learning (e.g., Gullberg, 2006; Taleghani-Nikazm, 2008). In what follows, we review the literature showing the role played by both referential (i.e., gestures that depict their referents, such as metaphoric and iconic gestures)¹ and nonreferential gestures in L2 learning from various perspectives, ranging from lexical learning to phonological learning. The present study will explore the potential role of observing and producing a type of visuospatial gesture that mimics durational properties of speech on perceiving and producing non-native phonological contrasts.

BENEFITS OF GESTURES IN L2 VOCABULARY LEARNING

Recent research has shown the beneficial effects of hand gestures on L2 vocabulary learning (e.g., Allen, 1995; Kelly et al., 2009; Macedonia & Klimesch, 2014; Tellier, 2008; see Macedonia, 2019 for a review). Allen (1995) trained 112 English participants to learn 10 French idiomatic expressions either with emblematic gestures or without gestures. Her results showed that training with gestures led to greater immediate recall and a smaller decay in recall after two months than training without gestures. Along the same lines, Tellier (2008) taught 20 L1 French children eight novel English words, by observing pictures related to the meaning of the target words or by observing and producing iconic gestures. The results revealed that training with gestures helped children to recall more words than training with pictures. Later, Kelly et al. (2009) trained 28 adult English-speakers to learn 12 Japanese verbs under four conditions: speech, speech + congruent iconic gesture, speech + incongruent iconic gesture, and repeated speech. After training, participants recalled the largest number of words under the speech + congruent gesture condition. In a 14-month classroom study, Macedonia and Klimesch (2014) investigated whether using iconic and metaphoric gestures helped L2 lexical learning. They taught 36 nonwords in an artificial language conforming to Italian phonotactics to 29 native German-speakers. Participants learned more words by performing gestures than by only repeating the words, showing that performing gestures significantly enhanced vocabulary learning in the long term (both 73 and 444 days after training).

Although most of the research on the role of gestures in L2 contexts focuses on the role of iconic and metaphoric gestures, recent evidence has shown that beat gestures may also be important in the acquisition of L2 vocabulary and pronunciation.

BENEFITS OF BEAT GESTURES IN L2 VOCABULARY AND PRONUNCIATION LEARNING

Beat gestures are a type of nonreferential hand gesture that are typically associated with prosodic prominence in speech and function as highlighters of rhythm (e.g., McNeill, 1992; Prieto et al., 2018). Several experimental studies have shown evidence of the beneficial role of using beat gestures for the learning of L2 vocabulary and pronunciation (e.g., Gluhareva & Prieto, 2017; Kushch et al., 2018). In a within-subject study, Kushch et al. (2018) trained 96 Catalan participants to remember 16 Russian new words presented with (a) prosodic prominence only; (b) beat gestures only; (c) both prosodic prominence and beat gestures; or (d) no cues. They found that target words presented with the combination of gestural and prosodic cues to prominence revealed the strongest learning effects. To assess the effects of beat gestures in L2 pronunciation learning, Gluhareva and Prieto (2017) trained 20 Catalan learners of English with videos in which an English instructor gave spontaneous responses to discourse prompts, either accompanied with beat gestures or not. Participants' own answers to the prompts were recorded before and after training and evaluated for accentedness. The results showed that observing beat gestures improved participants' pronunciation of the more difficult items. In similar studies, clapping hands to the rhythm of words has also been found helpful in improving L2 pronunciation (Baills et al., 2018; Zhang et al., 2018).

PITCH GESTURES AND THE LEARNING OF L2 PITCH FEATURES

A considerable body of research has demonstrated that the use of pitch gestures (e.g., hand gestures mimicking *F0* contour) significantly improved the recall of words in tonal languages (Morett & Chang, 2015), as well as the perception (Hannah et al., 2016) and learning of L2 lexical tones (Baills et al., 2019). Morett and Chang (2015) taught 57 English-speakers 20 novel Mandarin words accompanied by (a) "pitch gestures" to show the pitch information, (b) "semantic gestures" to show the words' meaning, or (c) unaccompanied by gestures. The results showed that pitch gestures helped the learners to memorize the Mandarin words differing in tone, suggesting that pitch gestures can strengthen the relationship between lexical meaning and tones. Later, Hannah et al. (2016) asked native English- and Mandarin-speakers to identify the Mandarin tones with or without gestural input. While the Mandarin-speakers performed at ceiling level, the English-speakers obtained significantly better scores with gestural input than without it, suggesting that gestural information lends a hand to the perception of novel tones. In a recent study, Baills et al. (2019) confirmed the benefits of observing and producing pitch gestures on the learning of Mandarin tones. In two experiments, they taught 18 minimal pairs of Mandarin words contrasting only in lexical tones to 106 Catalan-speakers by training them to either observe pitch gestures or both observe and produce the gestures. The results revealed that both observing and producing pitch gestures favored the learning of L2 tonal patterns and vocabulary.

The benefits of pitch gestures have also been shown at the sentence level. Kelly et al. (2017) reported that gestures signaling the sentence-final pitch features of Japanese yes/no questions and affirmative questions helped listeners to identify intonational distinctions. In line with this study, Yuan et al. (2019) confirmed the beneficial effects of observing pitch gestures in the learning of L2 intonation. They trained 64 Mandarin-speakers with

basic Spanish proficiency to learn three common Spanish intonation patterns (e.g., those for statements, yes-no questions, and requests) by either observing speech or observing speech with pitch gestures that represented nuclear intonation contours. Their results showed that training with gestures improved participants' realization of the intonation patterns in speech production more than training without gestures, suggesting that observing pitch gestures can favor the learning of L2 intonational patterns.

DURATIONAL GESTURES AND THE LEARNING OF L2 VOWEL-LENGTH CONTRASTS

In contrast with the positive role played by beat gestures and pitch gestures in the acquisition of L2 prosodic patterns, recent studies using durational gestures on the perceptual processing of Japanese durational vowel contrasts have yielded mixed results (e.g., Hirata & Kelly, 2010; Hirata et al., 2014; Kelly et al., 2014, 2017). Hirata and Kelly (2010) reported that while observing lip movements had positive effects on the acquisition of Japanese vowel-length contrasts, observing the gestures employed in their experiment (a beat gesture representing the short vowel and a hand sweep for the long vowel) did not show this effect. Later, Hirata et al. (2014) compared the effects of observing and producing two types of gestures representing length, namely syllable gestures (a hand sweeping representing a long vowel and a beat gesture representing a short vowel) and mora gestures (two beat gestures for a long vowel and one beat gesture for a short vowel) on auditory learning of vowel-length contrasts in Japanese. However, all the training methods were found to have similar effects on learning. In a follow-up study, Kelly et al. (2014) found that neither syllable gestures nor mora gestures showed any positive effect on either auditory perception or lexical learning. Furthermore, Kelly et al. (2017) demonstrated that despite the positive effect of observing pitch gestures on the perception of L2 intonational patterns, observing hand gestures representing vowel-length contrasts (the same as those used in Hirata & Kelly, 2010) still did not help participants to hear differences in vowel length. Taken together, the experiments carried out by Hirata, Kelly, and colleagues suggest that neither observing nor producing gestures signaling vowel length facilitates the perception of durational contrasts. They thus claimed that while visuospatial gestures were useful in acquiring intonational contrasts, they had only limited effects on the perception of durational contrasts. They concluded that durational gestures, in contrast with pitch gestures, were "a visual metaphor of a subtle auditory distinction within a syllable at the segmental level" (Kelly et al., 2017, p. 8).

However, despite these conclusions, gestures continue to be used in educational contexts for the teaching and learning of L2 pronunciation features (e.g., Hudson, 2011; Roberge et al., 1996; see Smotrova, 2017 for a review). In the context of the Verbotonal Method, Roberge et al. (1996) proposed a series of gestures intended to facilitate the acquisition of L2 Japanese pronunciation, including durational contrasts, which was a horizontal hand-sweep gesture mimicking short and long vowels. Hudson (2011) analyzed a 10-hour classroom video recording and observed the intensive use of various gestures by the instructor. The instructor employed hand gestures to mark durational features, with both hands moved horizontally outward to represent long vowels, and thumbs and index fingers pressed together to represent short vowels. Though the previously mentioned gestures differ in terms of specific hand shapes, both of them map temporal duration onto a spatial movement.

In our view, the negative results obtained in some of the previously mentioned studies may have been due to methodological reasons. The use of the contrasting pair of beat gesture and sweeping gesture as mimicking short and long vowel distinctions might not be effective for learners. Specifically, the use of a beat gesture for a weak short syllable is partially contradictory with its nature as a visual prominence indicator. The authors admit that they “may have chosen a wrong type of gesture to distinguish long and short vowels in language perception” (Hirata & Kelly, 2010, p. 305). Following up on observations by Roberge et al. (1996) and Hudson (2011), we believe that the use of a horizontal hand-sweep gesture of different durations (the longer the vowel, the farther the hand movement) might be more effective to mimic a vowel-length difference in space.

Importantly, there is behavioral evidence linking visual horizontal movements with the mental representation of duration. Casasanto and Boroditsky (2008) reported a series of six experiments in which participants viewed 162 horizontally growing lines on a screen and then replicated either their duration or their displacement by clicking or drawing with a mouse on a computer screen. These lines varied in duration (1–5 seconds in half-second increments) and displacement rate (200–800 pixels in 75-pixel increments). While in Experiment 1, participants had to replicate either duration or displacement without knowing the task until after the stimulus line had disappeared. By contrast, in Experiment 2 they were told which domain (i.e., duration or spatial displacement) they would have to replicate before each trial. The results showed that in both experiments, the spatial displacement of the moving stimulus strongly modulated people’s estimation of duration; however, reproducing the spatial displacement was not affected by duration, regardless of whether they were instructed to pay selective attention to a specific domain. Importantly, these results did not change even when extra information, like a constant temporal frame of reference (Experiment 3) or concurrent tone accompanying each growing line (Experiment 4), was provided; or when the growing line was replaced by a moving dot (Experiment 5) or a stationary line (Experiment 6). These consistent results suggest that the perception of durational contrasts in speech should be facilitated by contrasting horizontal movements that can be produced by the hands.

GOAL OF THE STUDY

The present study examined the effects of a horizontal sweep hand gesture encoding durational differences on the perception and production of Japanese words contrasting in vowel length by Catalan-speakers without knowledge of Japanese. Japanese has five vowels, /a/, /e/, /i/, /o/, and /u/, all which have durational contrasts (short and long) that can distinguish word meaning (e.g., *ike* “pond” vs. *ike:* “reverence”). By contrast, Central Catalan has seven vowels /a/, /e/, /e/, /i/, /o/, /ɔ/, and /u/, but none of them shows durational contrast (Wheeler, 2005). This study thus expands on preceding investigations by assessing the role of hand gestures encoding durational contrasts not only in *perception* but also in *production*. Because Catalan makes no phonemic distinctions based on vowel length, we hypothesize that visuospatial cues in the form of hand gestures mimicking vowel length might help Catalan-speakers without any knowledge of Japanese to perceive and to produce vowel-length contrasts. First, in relation to perception, training Catalan-speakers in the observation of durational hand gestures might enhance their accuracy in identifying Japanese vowel-length contrasts. Second, with regard to production, training

participants to actively produce durational hand gestures while producing the Japanese vowel-length contrasts might help them to better approximate a nativelike ratio of long to short vowel durations in Japanese speech production.

INDIVIDUAL DIFFERENCES AND L2 PRONUNCIATION

Apart from the effects of training, individual differences were found to have a considerable effect on pronunciation learning. For instance, listeners' musical experience and music perception abilities can strongly influence the learning of various pronunciation features (for a review, see Chobert & Besson, 2013). First, regarding the role of musical experience and musicianship, it has been found that musicianship boosts the learning of tonal languages (Cooper & Wang, 2012) because musicians are more sensitive to subtle changes in linguistic pitch than nonmusicians (Martínez-Montes et al., 2013). Musical experience has also been shown to enhance listeners' sensitivity to rhythm in a second language (Boll-Avetisyan et al., 2016). Furthermore, as Sadakata and Sekiyama (2011) suggested, "musicians may enjoy an advantage in the perception of acoustical features that are important in both language and music, such as pitch and timing" (p. 1). Second, in relation to music perception abilities, learners' perceptual abilities of nonlexical pitch patterns strongly correlate with the learning of lexical pitch patterns (Li & DeKeyser, 2017; Wong & Perrachione, 2007). Pitch-specific perception measures were also found to be the best predictor of successful learning of lexical tones (Bowles et al., 2016) and intonation analysis skills (Dankovičová et al., 2007).

Also, working memory capacities have been found to be relevant not only for L2 learning of vocabulary or grammar but also for L2 pronunciation learning (Juffs & Harrington, 2011, see Rota & Reiterer, 2009 for a review). Specifically, greater working memory capacities correlate with (a) better L2 narrative development (O'Brien et al., 2006), (b) greater fluency, complexity, and accuracy in L2 speech production and perception (Aliaga-García et al., 2010; Fortkamp, 2000), as well as (c) better inhibition patterns of the learners' L1, resulting in reduced negative transfer (Trude & Tokowicz, 2011). Working memory also predicts learners' speech outcome better than other factors such as imitation ability or attitude toward the area where the dialect is spoken (Baker, 2008).

The present study will thus assess the role of hand gestures encoding durational contrasts not only in L2 perception but also in L2 production processes. Importantly, we will control for the individual factors, namely musical experience, self-perceived musical skills (musicianship), music perception skills, and working memory abilities.

METHODS

The experiment consisted of a between-subjects training session with a pretest–posttest design, where participants were trained with 10 pairs of Japanese disyllabic words featuring vowel-length contrasts under one of two conditions: (a) either they watched two instructors pronouncing the words while performing gestures (the Gesture group, henceforth G group), (b) or they watched the same instructors pronouncing the same words without gestures (the No Gesture group, henceforth NG group). In both conditions, participants were asked to imitate the instructors, that is, to repeat the words in the NG group and to repeat the words and perform the gestures in the G group.

PARTICIPANTS

Fifty right-handed Catalan-speaking students (44 females, $M_{age} = 19.86$ years, age range: 18–29 years) were recruited from the Universitat Pompeu Fabra. Prior to the experiment, participants answered a questionnaire about their age, gender, linguistic background (percentage of dominance of Catalan relative to Spanish and foreign language ability), and musical background (number of years studying music, instruments played, amount of time spent on a regular basis listening to music and/or singing, and self-perceived music skills). All the participants reported speaking Catalan more than 75% of the time in daily verbal communication and none of them had studied Japanese before.

MATERIALS

This section describes the materials used in the familiarization phase, training session, pre- and posttests, and two control tasks, one to test music perception skills and the other to test working memory.

Audiovisual Materials for the Familiarization Phase

For the familiarization phase, a short 1.5-minute audiovisual sequence was created to introduce the Japanese vowel system, especially to illustrate the vowel length contrasts, and a brief description of the experiment.

Audiovisual Materials for the Training Phase

The training stimuli consisted of 10 pairs of Japanese disyllabic words contrasting in vowel length (see Table 1). Five pairs were unaccented with the LH(H) accentual pattern (e.g., *joko*² “side”), while the other five pairs were accented with the HL(L) pattern (e.g., *ító* “thread”). For all the words, the vowel-length contrasts were located in the word-final syllable (e.g., *joko* “side” vs. *joko:* “rehearsal”). This is because the word-final durational contrast has been found to be the most difficult for learners of Japanese to

TABLE 1. Ten minimal pairs of Japanese words and their English glosses for the training of vowel-length contrast

Word	Phonemic transcription ^a	English gloss	Word	Phonemic transcription ^a	English gloss
<i>joko</i>	<i>yoko</i>	side	<i>joko:</i>	<i>yoko:</i>	rehearsal
<i>cate</i>	<i>xare</i>	joke	<i>cate:</i>	<i>xare:</i>	reward
<i>kaze</i>	<i>kaze</i>	wind	<i>kaze:</i>	<i>kaze:</i>	taxation
<i>goke</i>	<i>goke</i>	widow	<i>goke:</i>	<i>goke:</i>	word form
<i>toko</i>	<i>toko</i>	bed	<i>toko:</i>	<i>toko:</i>	voyage
<i>ító</i>	<i>ito</i>	thread	<i>ító:</i>	<i>ito:</i>	east
<i>dzíko</i>	<i>tgíko</i>	accident	<i>dzíko:</i>	<i>tgíko:</i>	affairs
<i>kúco</i>	<i>kuro</i>	black	<i>kúco:</i>	<i>kuro:</i>	troubles
<i>kádo</i>	<i>kado</i>	corner	<i>kádo:</i>	<i>kado:</i>	art of poetry
<i>ído</i>	<i>ido</i>	water well	<i>ído:</i>	<i>ido:</i>	medicine

^aThe phonemic transcription conformed to Catalan orthography to facilitate reading by Catalan-speakers.



FIGURE 1. Screenshots of two trials of the training session in NG condition (upper panel) and in G condition (lower panel). In the G condition, the male instructor is showing the gesture produced while pronouncing the short vowel and the female instructor is showing the gesture produced while pronouncing the long vowel.

perceive (Tajima et al., 2008). All the syllables in the target words complied with the phonotactic constraints of Catalan.

Two right-handed native-speaking Japanese instructors (one female) were videotaped while producing the target word pairs. A total of 80 video clips were recorded (10 pairs of words \times 2 length contrasts \times 2 conditions \times 2 instructors). All video recordings were performed in a professional video-recording studio with a PDM660 Marantz professional portable digital video recorder and a Rode NTG2 condenser microphone. The videos featured a white background, and the upper half of the instructors' bodies and their faces were deliberately not blurred so that both groups had access to face and lip information.

Prior to recording, the two instructors received brief training on how to perform speech and gestures in accordance with our research needs. For the NG condition, both instructors produced the target pairs of words in a natural way and without moving any part of their body apart from their lips. For the G condition, they spoke the same set of target words while making the stipulated hand gestures: both instructors were asked to place their right hand in front of their body with the palm facing the floor and then produce a horizontal palm-down gesture to the right side synchronized with the duration of the target vowels (as illustrated in Figure 1). The durational contrasts were thus illustrated by the duration of the gesture; the longer the vowel, the longer the spatial movement. For each word, the instructors made a slight pause with the hand to indicate the syllabic boundary.

After recording, the videos were edited with Adobe Premiere Pro CC 2018 software. First, the videos were digitally flipped so that the movement appeared to be made with the left hand and participants could mirror the gestures with right hands. To control for any potential differences in the audio stimuli across the two conditions, the audio track recorded in the NG condition was added to the video track of the G condition, replacing the originally recorded audio material. To check that the resulting stimuli sounded natural, three Japanese native speakers assessed the naturalness of the videos with a 5-point Likert scale (1 = *very unnatural* and 5 = *very natural*). The results showed that the target stimuli sounded very natural ($M = 4.810$, $SD = 0.490$).

The training session consisted of the presentation of 10 pairs of words (block 1) followed by a repetition of these 10 pairs of words (block 2). Figure 1 visually illustrates the temporal sequence of presentation for two pairs of words as they appeared in each condition. For each pair, first, a black screen appeared with the phonemic transcription conformed to Catalan orthography of the two words always in the same order (the word with a short vowel followed by the word with a long vowel); second, a short video with one of the two instructors speaking the word with or without gestures (depending on the condition) was played; and finally, a 5-second black screen appeared, allowing participants to either repeat the word or repeat the word while imitating the hand gesture, depending on the condition. Five of the word pairs featured one instructor and the other five pairs featured the other instructor. The full sequence of 10 pairs was shown twice, with the pairs appearing in a different order the second time they were shown. However, the order of words in each pair did not vary (first short vowel, then long vowel).

Auditory Stimuli for the Pre- and Posttest Tasks

Vowel-length identification task. The auditory stimuli for the pre- and posttest vowel-length identification task consisted of four carrier sentences embedding 20 words featuring the vowel-length contrast in word-final position. Half of these words also appeared in the training session, and the other half did not.

The four carrier sentences each consisted of three sentence-initial syllables and three sentence-final syllables so that the target words always appeared in the central position (see Table 2). The reason for having various carrier sentences was to minimize fatigue caused by monotony. For each test, half of the sentences were uttered by one speaker and the other half by the other speaker.

The audio recordings were performed in a radio studio using professional equipment, and later edited with Audacity 2.1.2 software. All sentences were recorded twice at a

TABLE 2. Target word pairs and carrier sentences used in the pre- and posttest vowel-length identification tasks

Word pairs (English gloss)	Carrier sentences	
	Pretest	Posttest
<i>toko/toko</i> : (bed/voyage) <i>joko/joko</i> : (side/rehearsal) <i>kádo/kádo</i> : (corner/poetry art)	[M] <i>Kore-ga</i> ___ <i>to jomu.</i> “This is pronounced as ___”	[F] <i>Are-ga</i> ___ <i>dearu.</i> “That is ___”
<i>ídofído</i> : (water well/medicine) <i>ǎiko/ǎiko</i> : (accident/affairs)	[F] <i>Soko-wa</i> ___ <i>ga nai.</i> “There does not exist ___”	[M] <i>Soko-wa</i> ___ <i>ga aru.</i> “There exists ___”
<i>sotsu/sotsu</i> : (miss/communication) <i>ore/ore</i> : (I-masculine/thank)	[M] <i>Are-ga</i> ___ <i>dearu.</i> “That is ___”	[F] <i>Kore-ga</i> ___ <i>to jomu.</i> “This is pronounced as ___”
<i>mizo/mizo</i> : (ditch/unprecedented) <i>kíjo/kíjo</i> : (service/skillful) <i>rika/rika</i> : (science/liquor)	[F] <i>Soko-wa</i> ___ <i>ga aru.</i> “There exists ___”	[M] <i>Soko-wa</i> ___ <i>ga nai.</i> “There does not exist ___”

Note: [F] = female speaker; [M] = male speaker.

TABLE 3. Target word pairs and carrier sentences used in the pre- and posttest vowel-length imitation tasks

Word pairs (English gloss)	Carrier sentences	
	Pretest	Posttest
<i>eare/eave</i> : (joke/reward)	[M] <i>Kore-ga</i> ___ <i>deatu</i> .	[F] <i>Are-ga</i> ___ <i>to jomu</i> .
<i>kaze/kaze</i> : (wind/taxation)	“This is ___.”	“That is pronounced as ___.”
<i>goke/goke</i> : (widow/word form)		
<i>ító/ító</i> : (thread/to the east of)	[F] <i>Kore-ga</i> ___ <i>deatu</i> .	[M] <i>Are-ga</i> ___ <i>to jomu</i> .
<i>kúso/kúso</i> : (black/troubles)	“This is ___.”	“That is pronounced as ___.”
<i>sake/sake</i> : (wine/leftist)		
<i>iso/iso</i> : (beach/transference)		
<i>áse/áse</i> : (sweat/Mencius)		
<i>íco/íco</i> : (suicide note/clothes)	[M] <i>Kore-ga</i> ___ <i>deatu</i> .	[F] <i>Are-ga</i> ___ <i>to jomu</i> .
<i>sáju/sáju</i> : (hot water/left-right)	“This is ___.”	“That is pronounced as ___.”

Note: [F] = female speaker; [M] = male speaker.

normal speech rate by the same two instructors as in the training session. Later, the clearest and most natural-sounding samples were selected for the final audio files. In total, 40 audio files were created (20 sentences \times 2 tests).

Vowel-length imitation task. The auditory stimuli for the pre- and posttest vowel-length imitation task consisted of two carrier sentences, one for each test, embedding 20 words featuring the vowel-length contrast in word-final position. Half these words also appeared in the training session, and the other half did not. The words and carrier sentences were different from those used in the vowel-length identification task. However, like in the identification task, the two carrier sentences consisted of three sentence-initial syllables and three sentence-final syllables so that the target words always appeared in the central position (see Table 3). For each target word, participants listened to it embedded in the first sentence in the pretest uttered by one speaker and the second time in the second sentence in the posttest uttered by the other speaker.

The recording and material preparation procedures were the same as those followed for the identification task. All these materials were later submitted to SurveyGizmo,³ an online survey software, to create the experimental procedure.

Materials for the Control Tasks

Music perception skills. Participants undertook a perceptual music test for pitch and rhythm through two subsets of the Profile of Music Perception Skills (PROMS) test developed by Law and Zentner (2012). The rhythm and pitch tests were chosen because these two acoustic features are central in the phonological description of the target Japanese words used in the present investigation, which are characterized by contrasting

patterns of duration and pitch accentuation. Each subtest consisted of 18 randomized trials of varying difficulty where participants had to listen to a series of audio files. In the pitch test, for each trial, participants listened twice to the same pure tone, followed by a short interval and a comparison pure tone. The participants then had to indicate whether the comparison pure tone differed from the initial two. In the rhythm test, for each trial, the participants heard the same rhythmic sequence played twice with nonmelodic drumbeats, followed by a short interval and another rhythmic sequence. Again, their task was to indicate whether the third sequence had the same rhythm as the first two. In their responses, the participants could choose among five options: definitely different, probably different, I don't know, probably the same, and definitely the same.

Working memory. Working memory was assessed by the maximum number of words that the participants could remember after listening to various sequences of words in Catalan, which is an adaptation of a free recall word list memory task (Zhang et al., 2018). A total of 24 lists composed of commonly used Catalan words were selected as the test materials (see Appendix A). The lists contained several words ranging in number from four (minimum) to nine (maximum). There were four lists for each of the six ranges.

The words were read by a native Catalan-speaker and videotaped in a soundproof room. The resulting video was then edited using Adobe Premiere Pro CC 2018 software and cut into sections each containing only one string of words. This generated a set of 24 video segments that were embedded into a PowerPoint presentation.

PROCEDURE

The experiment proper started with a familiarization phase in which the participants watched a 1.5-minute video introducing Japanese vowel-length contrasts. This was followed by the pretest, which consisted of the vowel-length identification task and the vowel-length imitation task, each lasting 3 minutes. After pretest, the participant underwent the audiovisual training session, which lasted 2.5 minutes. This was followed by the posttests, consisting of the same tasks as the pretests, and, finally, the working memory test. A summary of the experimental procedure can be seen in [Figure 2](#).

The experimental procedure was carried out in a quiet room. Participants were tested individually and video-recorded during the experiment to ensure that they performed the tasks correctly. No feedback was provided during the entirety of the experiment.

Prior to the experiment, participants signed a consent form and answered a questionnaire about their age, gender, and linguistic and musical background, as noted in the preceding text. They also performed the two music perception skill tests of rhythm and pitch the day before the experiment. To control for potential differences between the two experimental groups, participants were assigned to one of the two training conditions in such a way that average scores of the two tests by group would be similar (for NG condition, $n = 25$, $M = 21.700$, $SD = 4.858$; for G condition, $n = 25$, $M = 21.100$, $SD = 4.474$).

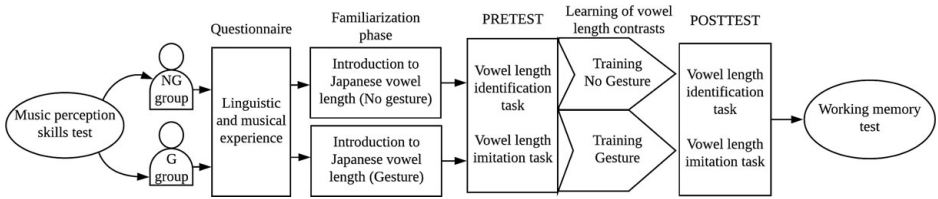


FIGURE 2. Experimental procedure.

Music Perception Skill Tests

The day before the experiment, participants were sent a link⁴ to access the rhythm and pitch tests online. Upon finishing the tests, their scores were automatically generated and exported from PROMS. The full procedure lasted approximately 15 minutes.

Familiarization Phase

In this phase, participants were familiarized with the Japanese vowel-length contrasts and the content of the training sessions depending on the group to which they were assigned. That is, participants in the NG group were shown how to repeat the words only, whereas participants in the G condition learned how to repeat the words while performing the gestures. The two contrasting words used in the familiarization phase were not included in the training phase that followed.

Pre- and Posttest Vowel-Length Identification Task

For this task, participants were instructed to work their way through a sequence of 20 online survey questions, each one appearing on a separate screen. Each screen offered written instructions in Catalan and a carrier sentence in Japanese written in Catalan-adapted phonemic transcription with a blank space in the middle (see the English translated screenshot in Figure 3 and list of carrier sentences in Table 3). A mouse click enabled participants to activate an audio recording to hear the sentence, which they were instructed to do only once per screen. Having heard the sentence, they clicked on a circle to indicate whether the second syllable of the target word had contained a long or a short vowel. Once they had done this, they proceeded to the next screen. The 20 audio items were automatically randomized by the software.

The target words, instructions, and procedure were the same for pretest and posttest. However, as noted in the preceding text, the order of carrier sentences and speakers varied across tests.

Pre- and Posttest Vowel-Length Imitation Task

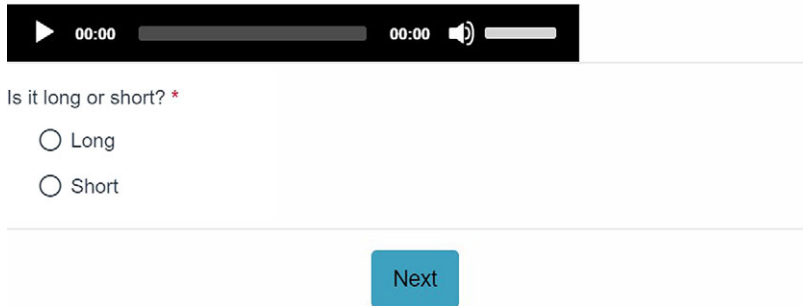
For the imitation task, participants worked their way through a continuation of the online survey, which in this case instructed them to repeat a total of 20 Japanese sentences with the target words embedded in the central position (see Table 3). However, the individual screens in this task merely showed written instructions—the carrier sentences were not

The sentence that you are going to hear is:

Ko re ga ___ to yo mu.

Click on "Play" to listen to the Japanese sentence. Pay attention to the last syllable of the word that appears in the center of the sentence (the blank space) and identify whether it is long or short.

Please listen to it ONLY ONCE.



Is it long or short? *

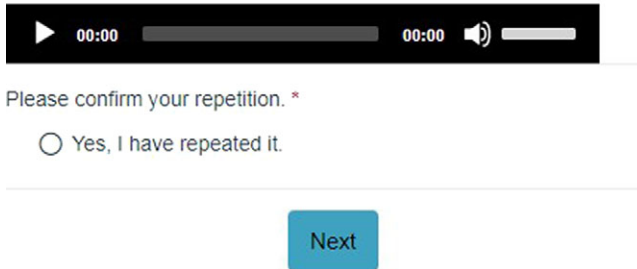
Long

Short

Next

FIGURE 3. Screenshot of a sample page from the vowel-length identification task (English translation).

Listen to the sentence ONLY ONCE. Then please repeat it.



Please confirm your repetition. *

Yes, I have repeated it.

Next

FIGURE 4. Screenshot of a sample page from the vowel-length imitation task (English translation).

presented in any written form (see the English translated screenshot in Figure 4). Here, after playing the audio file once, participants were supposed to repeat the sentence they had heard and then confirm that they had done so by clicking on a circle. Participants' oral production was recorded throughout the task. They then clicked on "Next" to move on to the next screen. Again, items were presented in a randomized order.

The target words and testing procedure were identical for pre- and posttest, except for the carrier sentences and speakers.

Training Phase

Participants watched the training video involving 10 pairs of words repeated in two blocks. In the NG condition, participants watched the instructor produce the word pairs consecutively and then repeated the words aloud. In the G condition, they watched the instructor produce the word pairs while performing the gestures and then repeated the words aloud while also mimicking the gestures. The training phase lasted approximately 2.5 minutes.

Working Memory Test

After having completed the posttest, each participant was assisted by the experimenters to complete the working memory test. This involved an experimenter taking the participant through a PowerPoint presentation in which were embedded short video files, each one featuring a list of words. Starting with the four-word strings, the participant first heard the list and then had to repeat it to the best of their ability. If the participant managed to repeat the full four-word list correctly, the experimenter moved on to the five-word strings, six-word strings, and so on. Whenever participants failed to repeat a string correctly, they were asked to move back to strings with a lower number of words. The final score equaled the maximum number of words in the lists that the participant could recall four times without errors.

CODING OF THE DATA⁵

Vowel-Length Identification Task

Participants' responses were assessed according to a binary rating system whereby a correct answer was given a score of "1" and an incorrect answer "0." The "Accuracy Rate" was obtained by calculating the percentage of correct answers over the total number of trials for each participant, with separate rates calculated for pretest and posttest.

Vowel-Length Imitation Task

To acoustically assess participants' performance on vowel-length contrasts, participants' oral productions during pre- and posttest were analyzed using PRAAT software (Boersma & Weenink, 2013). For each sentence, the initial and final boundaries of the target word and the final vowel of the target syllable were labeled. Thus, two tiers were created, a word tier and a target vowel tier (see Figure 5).

After annotation, the duration of each labeled vowel was automatically extracted by means of a PRAAT script.⁶ For each pair of words produced, a "Mean Duration Ratio" was calculated for each participant, with pretest and posttest ratios calculated separately. For each minimal pair in the same test, the Duration Ratio is equal to the duration of the long vowel divided by the duration of its short counterpart.

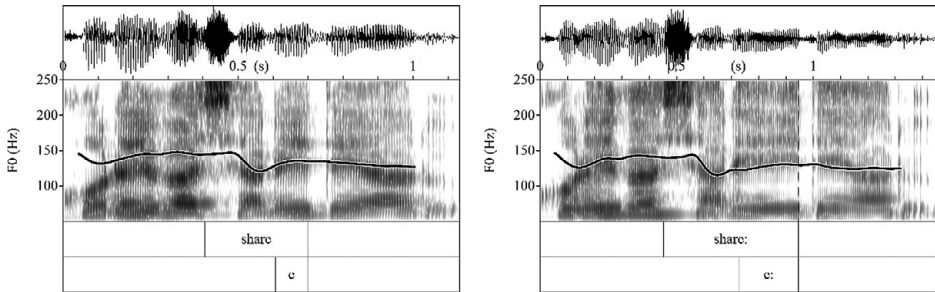


FIGURE 5. Spectrogram, pitch contour, and annotation scheme of the target Japanese word pair *share* “joke” (left panel) and *share:* “reward” (right panel) produced by a participant. The two tiers are the following: target words (“share” and “share:”) and starting and ending points of the target vowels (“e” and “e:”).

Musical Measures

The preexperimental questionnaire elicited information about each participant’s musical background (see Appendix B). Adapting Boll-Avetisyan et al.’s (2017) method, participants’ answers were coded as follows: (a) for the years spent studying music, one point for each year; (b) for the number of instruments played, one point for each instrument; and (c) for how often they reported singing and/or listening to music, 5 points if the participants had answered “daily” frequency, 4 points for “5–6 days per week,” 3 for “3–4 days per week,” 2 for “1–2 days per week,” 1 for “occasionally,” and 0 for “never.” These scores were then added to yield a “Musical Experience” variable. Following Law and Zentner (2012), the questionnaire also asked participants to characterize their self-perceived musical skills on a 5-point scale, ranging from 1 for “nonmusician” to 5 for “professional musician,” which was then labeled as “Self-Perceived Musical Skills.”

Regarding musical aptitude, participants’ scores on the music perception skill tests of pitch (labeled “Pitch Perceptual Ability”) and rhythm (labeled “Rhythm Perceptual Ability”) were automatically generated by the PROMS online testing system. To generate a categorical variable, a TwoStep Cluster analysis was applied using SPSS software in such a way that participants were automatically classified into two different levels in terms of Rhythm Perception Ability, namely higher ($n = 30$, $M = 28.100$, $SD = 2.936$) and lower ($n = 20$, $M = 18.733$, $SD = 4.042$). The same procedure was applied to classify the Pitch Perception Ability into two different levels, that is, higher ($n = 30$, $M = 23.967$, $SD = 3.057$), and lower ($n = 20$, $M = 14.850$, $SD = 3.407$). These two variables were used as independent variables, namely, “Rhythm Perception Level” and “Pitch Perception Level” in our models.

Working Memory

For each participant, the working memory score equaled to the number of words in the lists that the participant could recall four times without errors.

STATISTICAL ANALYSIS

The statistical analysis was carried out using IBM SPSS Statistics 24 (IBM Corporation, 2016).

First of all, we checked whether the participants in the NG and G groups were not statistically different in terms of Age, Musical Experience, Self-Perceived Musical Skills, Rhythm Perception Ability, Pitch Perception Ability, and Working Memory. Six independent samples *t*-tests were run and the results were as follows: (1) Age: $t(48) = -0.605$, $p = .548$; (2) Musical Experience: $t(48) = 0.034$, $p = .973$; (3) Self-Perceived Musical Skills: $t(48) = -0.241$, $p = .810$; (4) Pitch Perception Ability: $t(48) = 0.715$, $p = .478$; (5) Rhythm Perception Ability: $t(48) = 0.048$, $p = .962$; and (6) Working Memory: $t(48) = 0.215$, $p = .831$. These results confirmed that there was no significant difference between the two experimental groups.

For the vowel-length identification task, a Generalized Linear Mixed Model (henceforth GLMM) was run with Mean Accuracy Rate being the dependent variable. The fixed factors were Condition (two levels: NG and G), Test (two levels: pre- and posttest), and their interactions. Pitch Perception Level (two levels: higher and lower), Rhythm Perception Level (two levels: higher and lower), and Working Memory (scaled 4–7) were also included as fixed factors. Sequential Bonferroni comparisons were applied to the post-hoc pairwise comparisons.

For the vowel-length imitation task, a GLMM was run with Mean Duration Ratio being the dependent variable. The fixed factors were the same as in the GLMM applied to the vowel-length identification task.

In addition, in each task, the effect sizes (Cohen’s *d*; see Cohen, 1988) were calculated by comparing the means and standard deviations of the dependent variables at posttest and pretest.

RESULTS

VOWEL-LENGTH IDENTIFICATION TASK

Table 4 and Figure 6 show the mean Accuracy Rate obtained for the vowel-length identification task across conditions (NG and G) and tests (pretest and posttest). The descriptive data show that participants in the G group improved more (*Contrast estimate* = 8.000%) than those in the NG group (*Contrast estimate* = 4.200%) from pretest to posttest.

Table 5 summarizes the results of the GLMM analysis of the mean Accuracy Rate. The main effect of Test ($p < .001$) shows that participants’ Accuracy Rate differed significantly from pretest to posttest, and the main effect of Rhythm Perception Level ($p = .001$), suggests that participants’ rhythm perception ability is important for vowel-length

TABLE 4. Estimated mean, std. error, and 95% confidence interval for the accuracy rate (%) at pretest and posttest across conditions

Condition	Test	Estimated mean	Std. error	95% confidence interval	
				Lower	Upper
No gesture	Pretest	75.887	3.697	68.543	83.230
	Posttest	80.087	3.330	73.473	86.700
Gesture	Pretest	69.647	3.855	61.989	77.305
	Posttest	77.647	3.505	70.686	84.608

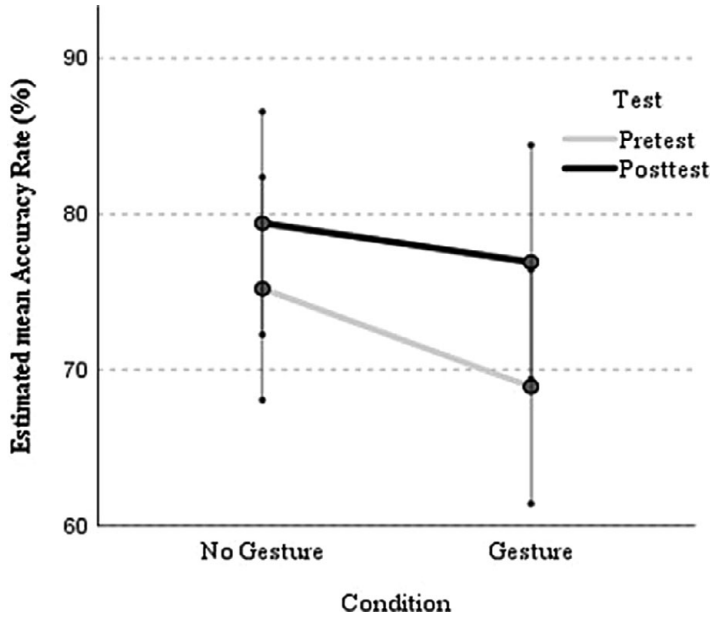


FIGURE 6. Estimated mean Accuracy Rates obtained in the vowel-length identification task across the Group (NG and G) and Test (pre- and posttest) conditions. Error bars indicate 95% CI.

TABLE 5. Summary of GLMM: Fixed effects for the mean accuracy rate of identification task

Fixed factors	<i>F</i>	<i>df1</i>	<i>df2</i>	<i>p</i>
Condition	1.945	1	91	.167
Test	15.851	1	91	<.001
Pitch perception level	0.020	1	91	.889
Rhythm perception level	15.511	1	91	.001
Working memory	0.126	3	91	.944
Condition × test	1.538	1	91	.218

identification. Post-hoc analyses revealed that participants obtained a significantly higher Accuracy Rate in the posttest than in the pretest (*Contrast estimate* = 6.100%; $t(91) = 3.981, p < .001$), confirming that participants improved significantly in vowel-length identification. Regarding the effect of Rhythm Perception Level, participants with higher Rhythm Perception Level obtained significantly higher Accuracy Rate than those with lower Rhythm Perception Level (*Contrast estimate* = 12.014%; $t(91) = 3.515, p = .001$), independently of the training condition or the test.

By contrast, no significant interaction between Condition × Test ($p = .218$) was found, suggesting that the improvement of the G group from pretest to posttest was not statistically larger than that of the NG group, although effect size for G group ($d = 0.594$) was larger than that for NG group ($d = 0.318$). In addition, Pitch Perception Level and Working Memory did not reveal any significant main effect.

TABLE 6. Estimated mean, std. error, and 95% confidence interval for the duration ratio at pretest and posttest across conditions

Condition	Test	Estimated mean	Std. error	95% confidence interval	
				Lower	Upper
No gesture	Pretest	1.641	0.137	1.370	1.912
	Posttest	1.938	0.137	1.667	2.209
Gesture	Pretest	1.511	0.143	1.226	1.796
	Posttest	2.517	0.143	2.232	2.802

IMITATION TASK

Table 6 and Figure 7 show the Mean Duration Ratio from the vowel-length imitation task across conditions (NG and G) and tests (pretest and posttest). The improvement in the Mean Duration Ratio from pretest to posttest for the G group (*Contrast estimate* = 1.006) was larger than that for the NG group (*Contrast estimate* = 0.297). Effect size was also larger for the G group ($d = 2.225$) than for the NG group ($d = 0.695$).

Table 7 illustrates the results of the GLMM analysis of the Mean Duration Ratio. These results revealed a main effect of Test ($p < .001$) and a significant two-way interaction of Condition \times Test ($p < .001$). Post-hoc comparisons revealed that participants improved significantly after training (*Contrast estimate* = 0.652, $t(91) = 14.862$, $p < .001$). Although the Mean Duration Ratio of the two groups did not statistically differ at pretest (*Contrast estimate* = 0.130, $t(91) = 1.010$, $p = .315$), the two groups obtained significantly different Mean Duration Ratios at posttest, with the G group outperforming the NG group (*Contrast estimate* = 0.578; $t(91) = 4.503$, $p < .001$). As for the control measures, that is, Rhythm Perception Level, Pitch Perception Level, and Working Memory, none of them showed significant main effect on the Mean Duration Ratio. These results suggest that although participants improved their duration ratio significantly after training, training with gestures led to a significantly larger improvement in the production task, regardless of the music perception skills and working memory capacities of the participants.

DISCUSSION AND CONCLUSION

The present study examined the effectiveness of visuospatial hand gestures depicting vowel-length features on perceiving and producing nonnative sounds. While previous studies have shown consistent beneficial effects of pitch gestures depicting pitch contour (e.g., Baills et al., 2019; Kelly et al., 2017; Morett & Chang, 2015; Yuan et al., 2019) and beat gestures representing rhythmic patterns (e.g., Gluhareva & Prieto, 2017), mixed results have been documented for the role of durational hand gestures, albeit tending toward the negative (Hirata & Kelly, 2010; Hirata et al., 2014; Kelly et al., 2014, 2017). Yet despite this lack of consistency, teachers frequently use a variety of visuospatial gestures to teach foreign language pronunciation, including durational contrasts (Hudson, 2011; Roberge et al., 1996). The present study further examined whether the use of durational hand gestures, produced with a horizontal hand sweep, is able to facilitate not only the perception but also the production of vowel-length contrasts

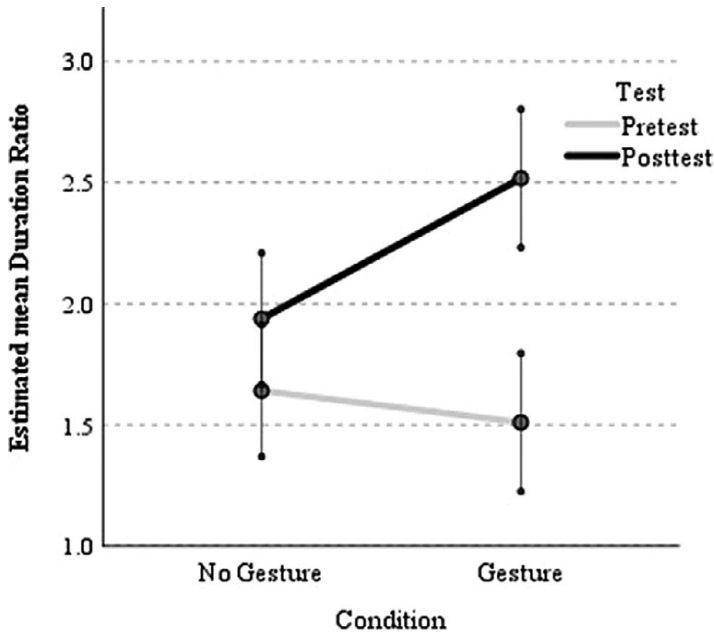


FIGURE 7. Estimated mean Duration Ratio obtained in the vowel-length imitation task across the Group (NG and G) and Test (pre- and posttest) conditions. Error bars indicate 95% CI.

TABLE 7. Summary of GLMM: Fixed effects for the mean duration ratio of imitation task

Fixed factors	<i>F</i>	<i>df1</i>	<i>df2</i>	<i>p</i>
Condition	3.451	1	91	.066
Test	220.864	1	91	<.001
Pitch perception level	0.117	1	91	.733
Rhythm perception level	2.714	1	91	.103
Working memory	0.449	3	91	.718
Condition × test	65.370	1	91	<.001

in Japanese. Following up on the results of Casasanto and Boroditsky’s (2008) psychophysical experiments showing that people’s estimation of duration could be modulated by spatial displacement, a proposal was made that using hand gestures that encode duration spatially (a horizontal sweep) might be effective for learning vowel-length contrasts in a second language.

The results of the identification task showed that participants improved significantly from pretest to posttest but training with gestures did not significantly enhance participants’ accuracy in perceiving the vowel-length contrasts in Japanese words more than training without gestures. Our findings are thus in line with the studies performed by Hirata, Kelly, and colleagues showing that either observing or producing durational hand gestures had limited effects in improving the *perception* of Japanese vowel-length contrasts.

However, previous studies did not assess the effects of durational hand gestures on production or pronunciation skills. The results of the imitation task showed that observing and producing durational hand gestures enhanced participants' accuracy levels in the pronunciation of vowel-length patterns as compared to training without gestures. The positive effects of gesture on production patterns found in the present study may be due to the visuospatial properties of the horizontal hand gestures used. In our view, this type of gesture encodes durational contrasts in speech in a more transparent way than the gestures used in previous studies. Recall that Hirata, Kelly, and colleagues used a beat movement encoding duration of a short vowel and a horizontal hand movement encoding duration of a long vowel. However, the association of a beat gesture with a target short vowel might be counterintuitive for speakers of languages like English where prominent syllables (e.g., longer and pitch accented syllables) are typically produced with beat gestures in spontaneous speech.

At first sight, it may seem surprising that observing and producing durational hand gestures had a positive effect at the productive level but not at the perceptual level. However, these asymmetric results might be related to the following reasons. First, as observed by Tajima et al. (2008), durational contrasts occurring in word-final positions in Japanese are harder for nonnatives to perceive than those occurring in other positions. In the identification task, participants started with a mean accuracy of 72.071% at pretest and ended up with a mean accuracy of 78.171% at posttest, revealing a small learning effect (less than 10%) after training. Moreover, while the perception task involved a challenging sentence-level identification task, the training just involved both perception and production of minimal word pairs presented in isolation. Therefore, a second reason for the asymmetric results might have been the role of carrier sentences, which may have triggered unequal difficulties and distractions across the two tasks. As noted in the preceding text, to avoid monotony, in the identification task the target words were embedded in four carrier sentences, but only two sentences were used in the vowel-length imitation task. Because in the identification task the carrier sentences varied considerably and were presented randomly, participants may have had trouble focusing their attention on the target words, thus diminishing the potential benefits of gestural input during training. However, because the imitation tasks featured a single carrier sentence at each test, participants could therefore more easily concentrate on the target words. A future study including a higher degree of consistency between training and tests might allow for a clearer assessment of the effects of producing and observing hand gestures on identifying durational contrasts. Finally, it might well be that when learning novel contrasting features, improvement in the perceptual dimension does not necessarily go hand-in-hand with improvement in the production dimension. In a longitudinal study, Nagle (2018) explored the long-term development of the L2 perception-production link in a pronunciation training course with 20 native English learners of Spanish. Participants had to learn the word-initial stops /b/ and /p/ in five sessions using 25 basic Spanish words. After each session, participants performed a sentence reading task and an identification task, both of which contained the trained words. The results showed that while participants improved significantly in both perception and production of /b/ and /p/ over the course of study, the performance in the reading task could not be predicted by the performance in the identification/perception task simultaneously in a single session. Our findings thus mirror those of Nagle's (2018) in relation to the lack of consistency

between L2 perception and production performance during pronunciation learning. In addition, other findings also support the lack of correlation between L2 speech perception performance and L2 production, suggesting that the two modules may be somewhat independent of each other (see Baese-Berk & Samuel, 2016; Zampini, 1998).

Regarding the relationship between the musical measures and L2 pronunciation learning, we found that rhythm perception skills positively affected participants' performance in the speech perception task. These results confirm previous findings suggesting that greater music perception skills may lead to better perception of durational variations in L2 speech (Sadakata & Sekiyama, 2011). Music perception skills may thus be an important individual factor to control for in future experiments on novel pronunciation learning (see Chobert & Besson, 2013). However, we could not find significant main effects of pitch perception skills in our speech perception task, perhaps due to the fact that the focus of the training task was on duration rather than pitch. Furthermore, music perception skills, either rhythm or pitch, did not have any significant main effect on speech production. This result is in line with previous studies that mainly showed correlations between perceptual abilities of music and language (e.g., Boll-Avetisyan et al., 2016; Cooper & Wang, 2012; Sadakata & Sekiyama, 2011; Wong & Perrachione, 2007).

In addition, working memory was not found to affect individual learning performance in either of the two tasks. Even though previous studies found working memory to be a good predictor of language learning (e.g., Rota & Reiterer, 2009), other studies have also claimed that working memory does not necessarily relate to the outcome of pronunciation learning (e.g., Mizera, 2006), nor does it predict learners' speech production better than other predictors (e.g., Posedel et al., 2012). Another reason for the lack of effect could be that we tested the working memory with real words in participants' L1 (Catalan), therefore, the influence of semantic meaning may have interacted with the participants' working memory performance. A future study might want to test whether working memory assessed with nonwords might increase its predictive status in pronunciation learning.

In sum, despite the null results on perception, our results show that durational hand gestures facilitate the pronunciation of novel words contrasting in vowel length. In the context of embodied learning, they provide clear empirical support for the view that multimodal trainings and self-performed gestures can help the learning of various aspects of nonnative pronunciation, especially at the suprasegmental level, and support recent practices in pronunciation teaching (e.g., Hudson, 2011; Smotrova, 2017). We believe that more experimental classroom studies are needed to further explore multimodal trainings for pronunciation teaching. All in all, the results of the study expand on recent studies that have highlighted the effectiveness of embodied instruction in second language learning by suggesting that gestures are a powerful tool that help learners to acquire not only vocabulary in second language (Macedonia, 2019) but also patterns of L2 pronunciation.

NOTES

¹According to McNeill's (1992) classification, co-speech gestures can be categorized as (a) iconic gestures, which employ spatial movements to mimic certain objects or actions; (b) metaphoric gestures, which represent abstract concepts; (c) deictic gestures, which are used to point toward a certain object or direction; and (d) beat gestures, which convey no referential meaning but seem to serve an accentuating function.

²The IPA transcription of Japanese used here follows Okada (1999). A mora with an accent marker is accented and carries a high pitch (in the current study, refers to the “HL(L)” pitch pattern) while a word with no accent marker begins with a low pitch and continues to be high pitched from the second mora onward (in this study, LH(H) pattern). (See Pierrehumbert & Beckman, 1988, pp. 7–8 for more details).

³<http://www.surveygizmo.com>

⁴<https://webapp.uibk.ac.at/psychologie/musiquote/index.php/656188/lang-ca-valencia>

⁵The data were coded by the first author.

⁶The script was created by Mietta Lennes and modified by Dan McCloy.

REFERENCES

- Aliaga-Garcia, C., Mora, J. C., & Cerviño-Povedano, E. (2010). Phonological short-term memory and L2 speech learning in adulthood. In K. Dziubalska-Kończak, M. Wrembel, and M. Kul (Eds.), *Proceedings of the 6th International Symposium on the Acquisition of Second Language Speech* (pp. 12–18). New Sounds. http://ifa.amu.edu.pl/newsounds/files/proceedings/proceedings_quotable_version.pdf
- Allen, L. (1995). The effects of emblematic gestures on the development and access of mental representations of French expressions. *The Modern Language Journal*, 79, 521–529.
- Baese-Berk, M. M., & Samuel, A. G. (2016). Listeners beware: Speech production may be bad for learning speech sounds. *Journal of Memory and Language*, 89, 23–36.
- Baills, F., Suárez-González, N., González-Fuente, S., & Prieto, P. (2019). Observing and producing pitch gestures facilitates the acquisition of Mandarin Chinese tones and words. *Studies in Second Language Acquisition*, 41, 33–58.
- Baills, F., Zhang, Y., & Prieto, P. (2018). Hand-clapping to the rhythm of newly learned words improves L2 pronunciation: Evidence from Catalan and Chinese learners of French. In K. Klessa, J. Bachan, A. Wagner, M. Karpiński, & D. Śledziński (Eds.), *Proceedings of the 9th International Conference on Speech Prosody* (pp. 853–857). International Speech Communication Association. <https://doi.org/10.21437/SpeechProsody.2018-172>.
- Baker, W. (2008). Social, experiential and psychological factors affecting L2 dialect acquisition. In M. Bowles, R. Foote, S. Perpiñán, & R. Bhatt (Eds.), *Selected proceedings of the 2007 Second Language Forum* (pp. 187–198). Cascadilla.
- Boersma, P., & Weenink, D. (2013). Praat: Doing phonetics by computer (version 6.0.28) [Computer software]. <http://www.praat.org>
- Boll-Avetisyan, N., Bhatara, A., & Höhle, B. (2017). Effects of musicality on the perception of rhythmic structure in speech. *Laboratory Phonology*, 8, 1–16.
- Boll-Avetisyan, N., Bhatara, A., Unger, A., Nazzi, T., & Höhle, B. (2016). Effects of experience with L2 and music on rhythmic grouping by French listeners. *Bilingualism: Language and Cognition*, 19, 971–986.
- Bowles, A. R., Chang, C. B., & Karuzis, V. P. (2016). Pitch ability as an aptitude for tone learning. *Language Learning*, 66, 774–808.
- Casasanto, D., & Boroditsky, L. (2008). Time in the mind: Using space to think about time. *Cognition*, 106, 579–593.
- Chobert, J., & Besson, M. (2013). Musical expertise and second language learning. *Brain Sciences*, 3, 923–940.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates.
- Cooper, A., & Wang, Y. (2012). The influence of linguistic and musical experience on Cantonese word learning. *Journal of Acoustical Society of America*, 131, 4756–4769.
- Dankovičová, J., House, J., Crooks, A., & Jones, K. (2007). The relationship between musical skills, music training, and intonation analysis skills. *Language and Speech*, 50, 177–225.
- Fortkamp, M. B. M. (2000). *Working memory capacity and L2 speech production: An exploration study* (Unpublished doctoral dissertation). Universidade Federal de Santa Catarina.
- Gluhareva, D., & Prieto, P. (2017). Training with rhythmic beat gestures benefits L2 pronunciation in discourse-demanding situations. *Language Teaching Research*, 21, 609–631.
- Goldin-Meadow, S. (2010). When gesture does and does not promote learning. *Language and Cognition*, 2, 1–19.
- Goldin-Meadow, S. (2011). Learning through gesture. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2, 595–607.

- Gullberg, M. (2006). Some reasons for studying gesture and second language acquisition (Hommage à Adam Kendon). *International Review of Applied Linguistics*, 44, 103–124.
- Hannah, B., Wang, Y., Jongman, A., & Sereno, J. A. (2016). Cross-modal association between auditory and visuospatial information in Mandarin tone perception. *The Journal of the Acoustical Society of America*, 140, 3225–3225.
- Hirata, Y., & Kelly, S. D. (2010). Effects of lips and hands on auditory learning of second-language speech sounds. *Journal of Speech, Language, and Hearing Research*, 53, 298–310.
- Hirata, Y., Kelly, S. D., Huang, J., & Manansala, M. (2014). Effects of hand gestures on auditory learning of second-language vowel length contrasts. *Journal of Speech, Language, and Hearing Research*, 57, 2090–2101.
- Hudson, N. (2011). Teacher gesture in a post-secondary English as a second language classroom: A sociocultural approach (*Unpublished doctoral dissertation*). University of Nevada.
- IBM Corporation. (2016). IBM SPSS statistics for Windows (version 24.0) [Computer software]. IBM Corporation.
- Juffs, A., & Harrington, M. (2011). Aspects of working memory in L2 learning. *Language Teaching*, 44, 137–166.
- Kelly, S. D., Bailey, A., & Hirata, Y. (2017). Metaphoric gestures facilitate perception of intonation more than length in auditory judgments of non-native phonemic contrasts. *Collabra: Psychology*, 3, 7.
- Kelly, S. D., Hirata, Y., Manansala, M., & Huang, J. (2014). Exploring the role of hand gestures in learning novel phoneme contrasts and vocabulary in a second language. *Frontiers in Psychology*, 5, 673.
- 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, 313–334.
- Kusch, O., Iguialada, A., & Prieto, P. (2018). Prominence in speech and gesture favour second language novel word learning. *Language, Cognition and Neuroscience*, 33, 992–1004.
- Law, L. N., & Zentner, M. (2012). Assessing musical abilities objectively: Construction and validation of the Profile of Music Perception Skills. *PLoS One*, 7, e52508.
- Li, M., & DeKeyser, R. (2017). Perception practice, production practice, and musical ability in L2 Mandarin tone-word learning. *Studies in Second Language Acquisition*, 39, 593–620.
- Macedonia, M. (2019). Embodied learning: Why at school the mind needs the body. *Frontiers in Psychology*, 10, 1–8.
- Macedonia, M., & Klimesch, W. (2014). Long-term effects of gestures on memory for foreign language words trained in the classroom. *Mind, Brain, and Education*, 8, 74–88.
- Martínez-Montes, E., Hernández-Pérez, H., Chobert, J., Morgado-Rodríguez, L., Suárez-Murias, C., Valdés-Sosa, P. A., & Besson, M. (2013). Musical expertise and foreign speech perception. *Frontiers in Systems Neuroscience*, 7, 1–11.
- McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago Press.
- Mizera, G. J. (2006). *Working memory and L2 oral fluency* (Unpublished doctoral dissertation). University of Pittsburgh.
- Morett, L. M., & Chang, L. Y. (2015). Emphasizing sound and meaning: Pitch gestures enhance Mandarin lexical tone acquisition. *Language, Cognition and Neuroscience*, 30, 347–353.
- Nagle, C. L. (2018). Examining the temporal structure of the perception–production link in second language acquisition: A longitudinal study. *Language Learning*, 68, 234–270.
- O'Brien, I., Segalowitz, N., Collentine, J., & Freed, B. (2006). Phonological memory and lexical, narrative, and grammatical skills in second language oral production by adult learners. *Applied Psycholinguistics*, 27, 377–402.
- Okada, H. (1999). Japanese. In International Phonetic Association (Ed.), *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet* (pp. 117–119). Cambridge University Press.
- Pierrehumbert, J., & Beckman, M. (1988). *Japanese tonal structure*. The MIT Press.
- Posedel, J., Emery, L., Souza, B., & Fountain, C. (2012). Pitch perception, working memory, and second-language phonological production. *Psychology of Music*, 40, 508–517.
- Prieto, P., Cravotta, A., Kusch, O., Rohrer, P. L., & Vilà-Giménez, I. (2018). Deconstructing beat gestures: A labelling proposal. In K. Klessa, J. Bachan, A. Wagner, M. Karpiński, & D. Śledziński. (Eds.) *Proceedings of the 9th International Conference on Speech Prosody* (pp. 201–205). International Speech Communication Association. <https://doi.org/10.21437/SpeechProsody.2018-41>.
- Roberge, C., Kimura, M., & Kawaguchi, Y. (Eds.). (1996). *Nihongo no hatsuon shidoo: VT-hoo no riron to jissai* [Pronunciation training for Japanese: Theory and practice of the VT method]. Bonjinsha.

- Rota, G., & Reiterer, S. (2009). Cognitive aspects of pronunciation talent. In G. Dogil & S. M. Reiterer (Eds.), *Language talent and brain activity* (pp. 67–96). Walter de Gruyter.
- Sadakata, M., & Sekiyama, K. (2011). Enhanced perception of various linguistic features by musicians: A cross-linguistic study. *Acta Psychologica, 138*, 1–10.
- Smotrova, T. (2017). Making pronunciation visible: Gesture in teaching pronunciation. *TESOL Quarterly, 51*, 59–89.
- Tajima, K., Kato, H., Rothwell, A., Akahane-Yamada, R., & Munhall, K. (2008). Training English listeners to perceive phonemic length contrasts in Japanese. *The Journal of the Acoustical Society of America, 123*, 397–413.
- Taleghani-Nikazm, C. (2008). Gestures in foreign language classrooms: An empirical analysis of their organization and function. In M. Bowles, R. Foote, S. Perpiñán, & R. Bhatt (Eds.), *Selected proceedings of the 2007 language research forum* (pp. 229–238). Cascadilla Proceedings Project.
- Tellier, M. (2008). The effect of gestures on second language memorisation by young children. *Gesture, 8*, 219–235.
- Trude, A. M., & Tokowicz, N. (2011). Negative transfer from Spanish and English to Portuguese pronunciation: The roles of inhibition and working memory. *Language Learning, 61*, 259–280.
- Wheeler, M. W. (2005). *The phonology of Catalan*. Oxford University Press.
- Wong, P. C. M., & Perrachione, T. K. (2007). Learning pitch patterns in lexical identification by native English-speaking adults. *Applied Psycholinguistics, 28*, 565–585.
- 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*, 5–32.
- Zampini, M. L. (1998). The relationship between the production and perception of L2 Spanish stops. *Texas Papers in Foreign Language Education, 3*, 85–100.
- Zhang, Y., Baills, F., & Prieto, P. (2018). Hand-clapping to the rhythm of newly learned words improves L2 pronunciation: Evidence from training Chinese adolescents with French words. *Language Teaching Research*. doi:10.1177/1362168818806531.

APPENDIX

APPENDIX A: WORD LISTS IN THE WORKING MEMORY TEST

Number of words	Catalan word strings
4	Coll, procés, govern, moviment Grup, festa, vila, silenci Família, raó, pell, escena Veritat, tipus, vi, producció
5	Rei, paraules, feina, llum, noia Silenci, consell, majoria, llit Llei, pedra, efecte, ciutat Cor, societat, realitat, favor, gent
6	Període, terme, origen, condicions, segle, punt Rei, boca, concepte, color, sang, acte, Coneixement, ciència, lloc, mar, teatre, joc Voluntat, posició, llocs, atenció, relacions, caràcter
7	Cos, quantitat, direcció, països, segles, acció, marit Cambra, unitat, guerra, consciència, posició, hores, punts Acord, importància, activitat, ombra, edat, imatge, carrer Peu, diners, qüestió, funció, moments, fusta, perill

Number of words	Catalan word strings
8	Muntanya, relació, església, foc, gust, existència, espai, paper Autor, sistema, flors, problema, pensament, llengua, vegada, situació Expressió, paraula, època, aigua, llei, pedra, efecte, ciutat Amor, moment, principi, aspecte, casos, veritat, tipus, vi, producció
9	Elements, canvi, pobles, lluna, aire, coll, procés, govern, moviment Grup, esperit, festa, història, vila, silenci, consell, majoria, llit Família, ànima, raó, població, llenguatge, experiència, banda, pell, escena Servei, fulles, nit, estudis, peus, idees, naturalesa, classe, vegades

APPENDIX B: LINGUISTIC AND MUSICAL BACKGROUND QUESTIONNAIRE (ENGLISH TRANSLATION)

Linguistic background

What percentage of Catalan do you use in your daily life?
 Apart from Catalan and Spanish, which language(s) do you speak?
 Have you ever studied Japanese?

Musical background

How many years of musical education have you ever received?

Do you play any instruments?

If yes, which instrument(s) do you play?

How often do you sing or listen to music?

- A. Every day
- B. 5–6 days per week
- C. 3–4 days per week
- D. 1–2 days per week
- E. Occasionally
- F. Never

Which one of the following best describes you?

- A. I'm a nonmusician
- B. I'm a music-loving nonmusician
- C. I'm an amateur musician
- D. I'm a semiprofessional musician
- E. I'm a professional musician