This is the accepted manuscript of the article. Please cite it in this form: 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

Hand Gestures Facilitate the Acquisition of Novel Phonemic Contrasts When They Appropriately Mimic Target Phonetic Features

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

Keywords: Hand gesture, segment learning, foreign language, aspiration contrasts, speech perception, speech production, second language pronunciation

Author Note

Xiaotong Xi, Department of Translation and Language Sciences, Universitat Pompeu Fabra; Peng Li, Department of Translation and Language Sciences, Universitat Pompeu Fabra; Florence Baills, Department of Translation and Language Sciences, Universitat Pompeu Fabra; Pilar Prieto, Institució Catalana de Recerca i Estudis Avançats and Department of Translation and Language Sciences, Universitat Pompeu Fabra.

This research was supported by funding from the Spanish Ministry of Science, Innovation and Universities (PGC2018-097007-B-I00) and the Generalitat de Catalunya projects (2017 SGR-971). The third author would like to acknowledge a predoctoral research grant awarded by the Department of Translation and Language Sciences, Universitat Pompeu Fabra.

No potential conflict of interest was reported by the authors. Correspondence concerning this article should be addressed to Xiaotong Xi, Department of Translation and Language Sciences, Universitat Pompeu Fabra, Roc Boronat 138, 08018 Barcelona, Spain. E-mail: xiaotong.xi@upf.edu

Abstract

Purpose: Research has shown that observing hand gestures mimicking pitch movements or rhythmic patterns can improve the learning of second language (L2) suprasegmental features. However, less is known about the effects of hand gestures on the learning of novel phonemic contrasts. This study examines (a) whether hand gestures mimicking phonetic features can boost L2 segment learning by naive learners and (b) whether a mismatch between the hand gesture form and the target phonetic feature influences the learning effect.

Method: Fifty Catalan native speakers undertook a short multimodal training session on two types of Mandarin Chinese consonants (plosives and affricates) in either of two conditions: Gesture and No Gesture. In the Gesture condition, a fist-to-open-hand gesture was used to mimic air burst, while the No Gesture condition included no such use of gestures. Crucially, while the hand gesture appropriately mimicked the air burst produced in plosives, this was not the case for affricates. Before and after training, participants were tested on two tasks, namely, the identification task and the imitation task. Participants' speech output was rated by five Chinese native speakers.

Results: The perception results showed that training with or without gestures yielded similar degrees of improvement for the identification of aspiration contrasts. By contrast, the production results showed that, while training without gestures did not help improve L2 pronunciation, training with gestures improved pronunciation, but only when the given gestures appropriately mimicked the phonetic properties they represented.

Conclusions: Results revealed that the efficacy of observing hand gestures on the learning of nonnative phonemes depends on the appropriateness of the form of those gestures relative to the target phonetic features. That is, hand gestures seem to be more useful when they appropriately mimic phonetic features.

Supplemental Material: https://doi.org/10.23641/asha. 13105442

Introduction

The acquisition of nonnative phonemes has been regarded as one of the major difficulties in foreign language pronunciation learning by adult learners. A growing number of studies have shown the positive role of multimodal perception training on the learning of novel phonemes (e.g., Hazan et al., 2005, among many others). This study aims at examining the effect of multimodal training with additional visual information, that is, hand gestures on both the perception and production of nonnative phoneme contrasts. Moreover, this study further examines how the appropriateness of the form of such hand gestures relative to the target phonetic features they are intended to represent will influence the training effect.

Multimodal Phonetic Training

Research in the last few decades has shown that language processing and language learning are multimodal in essence. Humans perceive speech through both auditory and visual information, and the two dimensions interact. First, speech perception can be influenced by visual input. The well-known McGurk effect (McGurk & MacDonald, 1976) has shown that speech perception may be strongly affected by visual information from speech articulators. Second, many studies have shown that the integration of auditory and visual information enhances listeners' speech perception in both noisy (Grant & Seitz, 2000) and normal (Arnold & Hill, 2001; also see Summerfield, 1992, for a review) environments and can improve speech recognition for listeners with hearing impairment (Grant et al., 1998) and nonnative listeners (Navarra & Soto-Faraco, 2007).

In the realm of second language (L2) acquisition, research has shown that the perception of nonnative phonemic contrasts is enhanced by the use of multimodal training paradigms integrating auditory and visual information as compared with auditory input alone (e.g., Hazan et al., 2005; Hirata & Kelly, 2010; among others). Hazan et al. (2005) tested the accuracy in

the perception of /v/–/b/–/p/ labial/ labiodental contrasts by 39 Japanese learners of English before and after a 4-week training session. Eighteen of the participants were trained with audiovisual information, which allowed them to listen to the instructor's speech and watch his or her lip movements at the same time. The remaining 21 participants were trained only with auditory information (i.e., they only had access to the instructor's speech). The results of a consonantal identification task after training indicated that audiovisual training was more effective in improving the perception of the labial/labiodental contrast compared to auditory training alone. Similarly, Hirata and Kelly's (2010) study showed that audiovisual training with access to the lip information led to a greater improvement than auditory-only training for 60 English speakers learning to perceive short and long vowels in Japanese. Importantly, success with this kind of audiovisual training seems to depend upon whether the articulatory movements of the L2 phoneme contrasts involve visually accessible information, such as lip movements. For example, for Japanese learners of English, the effect of audiovisual training reported in the abovementioned study by Hazan et al. was stronger for the labial/labiodental contrast compared to the /r/–/l/ contrast, which involves less visually salient features.

Regarding training effects on pronunciation, Hazan et al. (2005) found that perceptual training integrating both auditory and visual information had a beneficial effect not only on the perception of L2 phonemic contrasts but also on the pronunciation. This was confirmed by Inceoglu's (2016) study, in which 60 English learners of French were trained to learn three French nasal vowels. Participants were divided into three different modalities: audiovisual training, audio-only training, or no training. Participants' production of the target vowels was tested with an imitation task before and after training. The results showed that audiovisual training led to a significantly greater improvement in the production of the French nasal vowels than in the other two modalities, suggesting that multimodal training integrating visual and auditory information can facilitate pronunciation learning of nonnative phonemic contrasts.

Multimodal Phonetic Training with Hand Gestures

In L2 classroom studies, researchers have observed that teachers use a variety of gestures to help students achieve L2 target segmental and suprasegmental features of pronunciation such as syllabification, word stress, speech rhythm, and difficult target segments, among others (e.g., Smotrova, 2017; Zhang, 2002). Moreover, the effects of hand gestures on L2 pronunciation learning have also been assessed by a growing body of experimental research in laboratory settings, with a focus on the perception and production of suprasegmental features. For example, pitch gestures (i.e., hand gestures that mimic pitch contours in space) have been shown to enhance L2 lexical tone perception (Baills et al., 2019; Morett & Chang, 2015) and L2 intonation production (Yuan et al., 2019); beat gestures (i.e., hand gestures highlighting prosodic prominence through up-and-down rhythmic movements) have been shown to help reduce accentedness of L2 speech (Gluhareva & Prieto, 2017; Kushch, 2018; Llanes-Coromina et al., 2018). However, studies have shown contradictory effects of hand gestures on the perception and production of vowel length contrasts in Japanese (e.g., Hirata & Kelly, 2010; Hirata et al., 2014; Kelly & Hirata, 2017; Li et al., 2020). Some studies reported that hand gestures showed limited effects on improving the perception of durational contrasts (e.g., Hirata & Kelly, 2010; Hirata et al., 2014; Kelly et al., 2017), but another study suggested a beneficial role of hand gestures in the pronunciation of words carrying durational contrasts (Li et al., 2020). While most of the aforementioned studies demonstrated the positive role of hand gestures for the learning of L2 tones, intonation, and duration properties, as well as for improving L2 pronunciation in general, less is known about the potential effect of gestures at the segmental level.

Hand gestures can adopt a variety of forms that mimic phonetic information in speech and thus serve to highlight target phonetic properties at the segmental level, which has long been used by clinicians as a therapeutic approach to help the production of speech sounds. Klick

(1985) created the adapted cueing technique, which treated hand gestures as an additional visual input to enhance the speech produced by speakers with dyspraxia. A series of cospeech hand gestures were proposed to cue several characteristics of English sounds, such as the place of articulation, the manner of articulation, and the trajectory of the tongue, for example, hand gestures from the side of the face to the nose to mimic the tongue movement of nasal /n/; the quick and slow hand movements could mark the manner difference between stops and continuants. Shelton and Garves (1985) applied another therapeutic approach to the treatment of a child with developmental apraxia, that is, the signed target phoneme, which relied on the use of hand shapes from the American Manual Alphabet to cue the phonemes. The child's production accuracy of the speech sound /s/ improved after the treatment. The author mentioned that the visual stimuli of hand gestures may help the child to recall the association of sound and symbol, which contributed to the success of this therapy. Despite the widespread use of hand gestures in clinical practice, there is limited empirical evidence of the effectiveness of gesture-based treatment. In a recent study, Rusiewicz and Rivera (2017) conducted a single-subject experiment in which an adult patient diagnosed with apraxia of speech was trained to pronounce the /r/ sound in five vowel + /r/ contexts at the syllable level (/ɛr/, /ɔr/, /ɪr/, /ɑr/, and /ɑɪr/) with the use of a hand gesture mimicking the articulation of the /r/ sound. The perceptual ratings by 28 naive listeners revealed a significant improvement in the production accuracy of the /r/ sound by the patient, which suggested that hand gestures mimicking articulation could be used as a treatment strategy to help persons with apraxia better produce speech sounds.

In the context of L2 pronunciation learning, to our knowledge, only two empirical studies have assessed the role of the hand gestures mimicking phonetic features in L2 segmental learning (Amand & Touhami, 2016; Hoetjes et al., 2019). Amand and Touhami (2016) trained 16 French learners of English on the pronunciation of released and unreleased English stops /p/, /t/, and /k/ either with (Gesture [G] group) or without (No Gesture [NG] group) hand

gestures. For the G group, a fist-to-open-palm hand gesture was used to illustrate the released stops, while a stretched-fingers-to-fist gesture was used for the unreleased stops. Learners' production was tested by means of a reading task, which was undertaken before and after training. An acoustic analysis of the target stop consonants in the reading task revealed that training with gestures mimicking phonetic features yielded a significant improvement in the pronunciation of L2 English released and unreleased stops.

Recently, Hoetjes et al. (2019) have assessed the effects of pointing gestures (pointing a finger toward certain objects, e.g., toward the mouth in this study) and hand gestures mimicking articulatory information on the pronunciation of L2 Spanish phonemes /θ/ and /u/ by 51 Dutch native speakers. As for the hand gestures mimicking articulatory information, fingers were extended forward to indicate the tongue position for the phoneme /θ/ and the palm and the fingers formed an "o" shape in order to indicate the rounding of the lips for the /u/. Participants' pronunciation performance was tested by a sentence-reading task before and after training. Then, their production was assessed by two phoneticians and 46 native Spanish speakers successively. The results showed that observing pointing gestures had a beneficial effect on the learning of the two phonemes, while observing gestures mimicking articulatory information helped the learning of /u/ but impaired the learning of /θ/, which suggested that different types of gestures may have different effects depending on the L2 phoneme to be learned. The gesture used for the /θ/ consonant involves using the fingers to represent the shape of the tongue when this sound is being produced. However, in our view, this gesture failed to convey other important information about how to articulate this consonant, such as the fact that the tip of the tongue should be placed between the teeth and the jaw should be in a fairly open position. This suggests that the form of the hand gestures should appropriately mimic the intended target phonetic features; in other words, the hand gestures should match the phonetic properties of the target segments.

The importance of the correspondence between the form of the gesture and the target linguistic property it is representing has been assessed by several studies. First, speech perception may be influenced according to the matching/ mismatching of the gesture form (Hannah et al., 2017; Kelly et al., 2017). Hannah et al. (2017) found that pitch gestures could facilitate L2 lexical tone perception when the gestures matched the properties of the lexical tones (e.g., a rising gesture for a rising tone), whereas a mismatch between gesture form and pitch movement (e.g., a rising gesture for a falling tone) disrupted perception. Similarly, Kelly et al. (2017) found that congruous pitch gestures facilitated the perception of intonation in L2 Japanese and that congruous hand gestures mimicking durational features tended to help the processing of vowel length contrast. Second, the use of appropriate hand gesture could boost L2 lexical learning (e.g., Kelly et al., 2009; Macedonia et al., 2011) and L2 pronunciation learning (Zhen et al., 2019). For example, L2 words trained with congruent iconic gesture could lead to better word memorization, whereas a mismatching gesture did not facilitate L2 lexical learning (Kelly et al., 2009). Similarly, Zhen et al. (2019) showed that congruent pitch gesture could improve auditory perceptual learning of L2 lexical tones. Crucially, pitch gestures performed in the horizontal plane could still boost perceptual learning as long as they were congruent with pitch contours. All in all, while gesture form has been shown to influence the effect of training with hand gestures in L2 word and lexical tone learning, little is known about the role of gesture form in the context of L2 segment learning.

In view of the aforementioned findings, further research is needed to clarify the potential effect of hand gestures mimicking phonetic features on L2 segmental learning, as well as to assess how important it is that the form of the gesture be appropriate to the target phonetic features. We hypothesize that success with gesture trainings will depend upon whether the form of gesture appropriately represents visual or audible phonetic features of the L2 phonemes. Therefore, in this study, we investigated (a) whether hand gestures encoding phonetic features

can boost L2 segment learning by naive learners and, crucially, (b) whether the appropriateness of the hand gesture for the phonetic features in question has an effect on the learning outcome.

The Current Study

This study examined the effects of a hand burst gesture encoding aspiration property on the perception and production of Mandarin Chinese consonants contrasting in aspiration by Catalan speakers without knowledge of Mandarin. Mandarin has a total of six pairs of consonants, which have traditionally been described as contrasting in aspiration, and the acquisition of these aspirated phonemes is considered to be one of the main difficulties experienced by speakers of Romance languages when learning Mandarin (Chen et al., 2013). The three pairs of Mandarin plosives are voiceless unaspirated sounds (bilabial /p/, dental /t/, and velar /k/) with their aspirated counterparts (/p^h/, /t^h/, and /k^h/), while the three pairs of affricates are voiceless unaspirated sounds (alveolar /ts/, alveolopalatal /tɕ/, and retroflex /tʂ/) along with their aspirated counterparts (/ts^h/, /$tɕ^h$/, and /$tʂ^h$/; see Duanmu, 2007). Central Catalan phonological inventory has a set of plosives and affricates that do not match the Mandarin system. While the plosives are three voiced unaspirated sounds (bilabial /b/, dental /d/, and velar /g/) with their voiceless counterparts (/p/, /t/, and /k/), the affricates are two voiceless sounds (alveolar /ts/ and postalveolar /tʃ/) with their voiced counterparts (/dz/ and /dʒ/; see Wheeler, 2005). Thus, Catalan plosives and affricates contrast in voicing features, and it can be said that the aspiration feature present in Mandarin plosives and affricates does not exist in the Catalan sound counterparts. Therefore, both the perception and production of Mandarin aspirated consonants can be considered as difficult to learn for Catalan speakers.

Importantly, though Mandarin plosives and affricates have been traditionally described as having an “aspiration” contrast, studies have revealed that the aspirated plosives and affricates differ in their phonetic realization. From the point of view of phonetic articulation, plosives feature a complete occlusion of the vocal tract causing a buildup of an air pressure followed by

a sudden opening so that the air is let out quickly, which produces a perceivable burst of air. By contrast, though affricates feature a closure that is similar to that seen in plosives, the tract is opened gradually and compressed air is released gradually, which results in an audible frication (Laver, 1994). Thus, as can be observed in Figure 1, the two types of consonants differ in the strength of the air burst and the duration of the frication phase. Note that, in contrast with the affricates /tɕʰ/–/tɕ/ (bottom panels), the initial release of the plosives /tʰ/–/t/ (top panels) is characterized by a spike of acoustic energy visible in both the waveform and the spectrogram. After the opening, the aspirated consonants have a longer release time than their unaspirated counterparts, with affricates showing a more turbulent airstream than plosives. This is because, with unaspirated consonants, the vocal cords vibrate immediately after the air release, while aspirated consonants exhibit a longer release time, which allows more air to flow out and thus a greater drop in air pressure (Ladefoged, 1963).

Auditorily, aspirated plosives are characterized by a very salient audible air burst that is due to the changes in air pressure and the sudden release of air (see Figure 1, top panels). By contrast, aspirated affricates are not perceived to start with an air burst, as the longer release time allows the compressed air to be released over a longer frication period (see Figure 1, bottom panels).

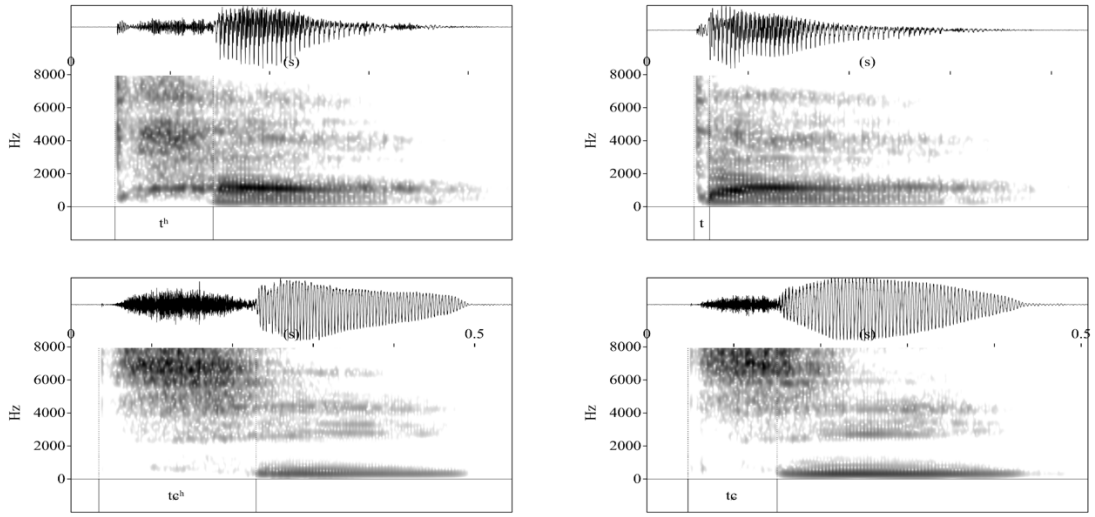


Figure 1. Waveform and spectrogram for the phonemic contrasts /t^h/–/t/ (top panels) and /$tɕ^h$/–/$tɕ$/ (bottom panels).

We adopted a fist-to-open-hand gesture (hand burst gesture, see Figure 2) from Amand and Touhami (2016) and Zhang (2002) to simulate the extra burst of air, which refers to the aspiration feature of Mandarin aspirated plosives as they are perceived with a stronger air burst and produced with more air release than the unaspirated plosives. We hypothesize that the use of the hand burst gesture will be beneficial for the learning of aspirated plosives. Crucially, this burst gesture closely mimics not only how an aspirated plosive is perceived (i.e., the prominent air burst) but also its phonetic articulation (i.e., the clenched fist illustrates the occlusion; and the suddenly opened palm, the quick opening). By contrast, this hand burst gesture is not congruent with the phonetic features of aspirated affricates (i.e., a longer frication that is easily perceivable), since the quickly opened hand is not a good visual representation of the more gradual air release presenting a longer frication. Thus, while the hand burst gesture closely matches the auditory and articulatory properties of aspirated plosives, this is not the case for aspirated affricates.

Figure 2. The hand burst gesture, in which clenched fists are rapidly opened, used in this study to depict the aspiration feature of Mandarin aspirated plosives. Image used with consent

from the instructor.

In summary, the goal of this study is twofold: first, to assess whether using a hand burst gesture that visually depicts the air burst and aspiration will facilitate the learning of nonnative aspiration contrasts and, second, to test whether the presence of a mismatch between the hand gesture and the target phonetic features influences phonological learning. Crucially, as we have noted, while the hand burst gesture used in the study closely matches the properties of aspirated plosives, this same gesture is a poor match for the properties of aspirated affricates.

Method

In a between-subjects study with a pre- and posttest design, participants were trained under one of two conditions to perceive and produce two types of consonants, plosives and affricates. In the G condition, they watched two instructors performing hand gestures when uttering the target words containing the aspirated consonants. In the NG condition, they watched the instructors utter the same words without producing any gestures.

As noted, the training items used in both conditions were Chinese words containing two types of consonants, which referred to a within-subject condition, the consonant type, with two levels: plosives and affricates. Since the hand burst gesture employed matched the phonetic properties of the plosives but not those of the affricates, the type of consonant represented the matching relation between gesture and phonetic realization in the G condition.

Participants

Fifty Catalan-dominant speakers ($N = 50$, 42 females and eight males; $M_{age} = 20.90$ years, $SD = 2.496$, $SE = 0.250$) were recruited from a public university in Barcelona. Prior to the experiment, participants were asked to answer a questionnaire about their age, gender, and linguistic and musical background (see Appendix). All the participants reported more than 75% use of Catalan in daily verbal communication, and none of them had studied Mandarin Chinese

before. They were then randomly assigned to one of the two training conditions, namely, NG (n = 25, 21 females and four males) and G (n = 25, 21 females and four males). They signed a written consent giving permission to process their data and received a small amount of money in compensation for their participation.

Materials

The experiment started with a familiarization phase, which consisted of a short introduction to the pronunciation of Mandarin Chinese aspirated consonants. Next, two pretest tasks (identification and imitation tasks) were followed by a short training session involving the learning of six pairs of Mandarin aspiration contrasts. The training session was followed by a posttest consisting of the same identification and imitation tasks performed in the pretest. This section describes the preparation of the materials used in the four phases of the experiment (familiarization, pretest, training session, and posttest; see Supplemental Materials S1–S5).

Audiovisual materials for the familiarization phase. For each condition, a short video featuring one of the instructors was created in order to introduce briefly the main features of the Mandarin aspiration contrasts and then describe the procedure that would be followed in the training sessions. The two contrasting words used as an example in the familiarization video (*chăng fáng* "factory" vs. *zhăng fáng* "eldest branch of a family") were not included in the subsequent training session, nor were they tested in any of the tasks. Participants in the G group observed the instructors producing this example word pairs with hand gestures, whereas participants in the NG group observed those same productions without any given gestures.

Audiovisual stimuli for the training session. Twelve disyllabic Chinese words containing the target phonemes were selected as training stimuli. The aspiration contrast was located in word-initial position for all the pairs (e.g., *pí yán* "dermatitis" vs. *bí yán* "rhinitis"). Half of the consonants in word-initial position were plosives, and the other half were affricates.

All the audiovisual materials were prepared in a professional broadcasting studio, with a PDM660 Marantz professional portable digital video recorder and a Rode NTG2 condenser microphone, and were edited with Adobe Premiere Pro CC 2018 software. Two native Mandarin instructors (one female and one male) were video-recorded as they produced the training stimuli for the two training conditions. The words with aspirated consonants were produced with (for the G group) and without (for the NG group) gestures by both speakers. The words with unaspirated consonants were not accompanied by any gestures in any of the conditions; in other words, the video clips of these items were the same across the two conditions. Thus, a total of 36 video clips were obtained (6 words with unaspirated consonants × 2 instructors + 6 words with aspirated consonants × 2 instructors × 2 conditions).

Before the recordings, the two instructors were trained to use the same hand burst gesture to accompany their production of aspirated consonants embedded in words: This involved simultaneously raising their two hands in clenched fists to a position somewhat higher than their shoulders and then opening them suddenly to splay their fingers with a forward movement toward the camera in order to illustrate the air burst (see Figure 2). Then, they immediately put back their hands in the rest position as they finished pronouncing the word.

To control for any potential differences in the auditory stimuli across the two conditions, the audio track recorded in the NG condition was copied onto the video track of the G condition, replacing the original audio track. In order to check whether the newly created videos sounded natural, three Mandarin native speakers assessed their naturalness with a 5-point Likert scale (1 = *very unnatural* and 5 = *very natural*). The results showed that the videos were perceived as natural by native speakers (M = 4.89, SD = 0.32, SE = 0.05). To get the participants progressively acquainted with the aspiration contrast, the training video consisted of two blocks that presented the minimal pairs in two different ways.

In the first block, for each minimal pair, the word containing the unaspirated consonant

was presented first, pronounced by both instructors, and followed by the word containing the aspirated counterpart, also pronounced by both instructors. For each word, the Catalan-adapted orthography of the words was first provided on the screen. Note that, in order to make the target sounds more visually salient, the symbols representing them were displayed in yellow, with the rest of the word in white. This was followed by the corresponding video clips of the target word being spoken by both instructors. This sequence was followed by a black screen. Two versions of the training videos were prepared to reflect the G versus NG condition. Thus, in the training video for the NG group, the instructors never performed any gestures, whereas in the video for the G group, the video clips for the words containing aspirated consonants showed the instructors accompanying those consonants with hand burst gestures (as noted, unaspirated consonants were not accompanied by gestures). In total, in the first block, the 12 words were presented twice (12 words × 2 instructors).

The second block was intended to train the words contrastively in pairs. First, the Catalan-adapted orthography of the minimal pair was presented on the screen. Then, each minimal pair was produced first by one instructor and then by the other. This sequence was followed by a black screen. The second block was repeated twice with different trial orders. Therefore, in total, each minimal pair was presented 4 times (1 word pair × 2 instructors × 2 times).

Auditory stimuli for the pre- and posttest identification task. For the identification task used in the pre- and posttest, six pairs of Mandarin disyllabic words featuring the aspiration contrast in word-initial position were chosen. Half of the words were included in the training session video, and the other half were not.

The audio recordings were performed in the radio studio using professional equipment and later edited with Audacity 2.1.2 software. All words were recorded twice at a normal speech rate by the same two instructors. Later, the clearest and most natural-sounding samples were selected for the final 24 audio files (12 words × 2 instructors). For the 12 words in both

pre- and posttests, half of the audio recordings were spoken by one instructor and the other half were spoken by the other instructor, with an obligatory alternation between pretest and posttest. That is, if in the pretest, the audio recording of a word was produced by one instructor, the recording of the same word for the posttest should be produced by the other instructor. The audio files were then uploaded to the online survey platform SurveyGizmo with the order of items automatically randomized within each test by the software.

Auditory stimuli for the pre- and posttest imitation task. The stimuli for the imitation task consisted of six pairs of Mandarin words, which were different from the words in the identification task. While half of these words appeared in the training materials, the other half did not.

The recording preparation procedure was the same as for the identification materials. As in the case of the identification task, at pretest, half of the stimuli were spoken by one instructor and the other half were spoken by the other instructor, and the gender of the speakers was counterbalanced at posttest.

Experimental Procedure

Participants were tested individually in a quiet room, and no feedback was provided during the entire experiment. Participants were video-recorded during the experiment to ensure that they performed the tasks correctly.

A summary of the experimental procedure can be seen in Figure 3. Prior to the experiment, participants signed a consent form and answered a questionnaire about their age, gender, and linguistic and musical background, as noted above. They were then randomly assigned to one of the two conditions, NG or G. The experiment started with each participant watching the familiarization video (3 min 11 s). Next, they performed the identification and imitation tasks making up the pretest. Immediately after the pretest, they watched the short training video (5

min 36 s). Finally, they completed the posttest, which consisted of the same tasks performed in the pretest. Altogether, the experiment lasted about 25 min.

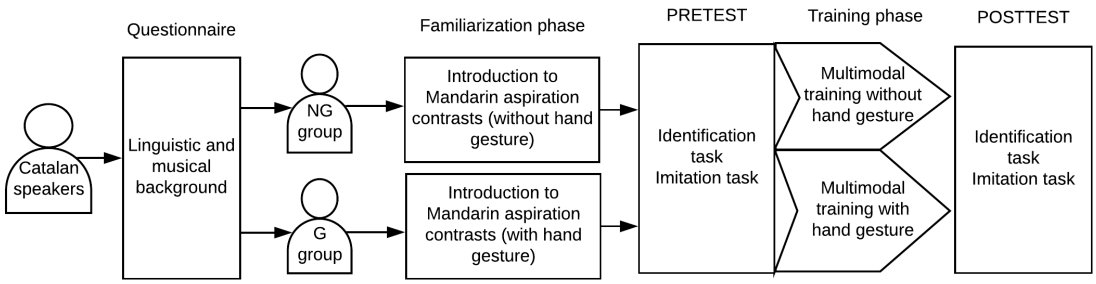


Figure 3. Experimental procedure. G = Gesture; NG = No Gesture.

Pre- and posttest identification task. For this task, participants were instructed to work their way through a sequence of 12 trials, each one appearing on a separate screen. Each screen offered written instructions in Catalan. A mouse click enabled participants to activate an audio recording to hear the target word, which they were instructed to do only once. Once they had heard the word, they clicked on a circle to indicate whether they had heard the word with an aspirated consonant or its unaspirated counterpart (the options were given in the adapted transcription). They then proceeded to the next screen.

Pre- and posttest imitation task. For the imitation task, participants were instructed to repeat a total number of 12 Chinese words. No transcription was given. After playing the audio file once, participants were asked to repeat the word they had heard and then to 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.

Training session. After the pretest, participants watched the training video involving six pairs of words featuring unaspirated and aspirated consonants. In the NG condition,

participants watched the instructors utter the words, whereas in the G condition, participants watched the instructors performing the gesture as they simultaneously produced the aspirated consonants. In both conditions, the participant was expected to remain motionless and silent as they watched the training video.

Data Coding

A total of 1,200 responses (50 participants × 12 identification questions × 2 tests) were obtained from the identification task at pre- and posttest, and a total of 1,200 recordings (50 participants × 12 imitation items × 2 tests) were obtained from the imitation task.

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 was given a score of "0."

Imitation task. The recordings obtained in the imitation task (a total of 1,200) were rated by five Mandarin native speakers (N = 5, three females and two males; M_{age} = 25.60 years, SD = 3.578, SE = 1.600), who were blind to the experiment. Raters were asked to assess two pronunciation variables from each audio file: (a) the general pronunciation accuracy of the target word and (b) the accuracy of the consonantal feature of the target consonants (i.e., aspiration for the plosives and duration of frication for the affricates). Every rater was asked to rate all the 1,200 recordings, thus obtaining a total of 6,000 rating scores for each measure.

All the recordings were presented to the raters randomly so that they were unaware of the training conditions and tests. Before rating, all raters were trained in a 30-min session with some examples so as to become familiar with the evaluation system. Raters first rated the general pronunciation of all audio recordings and then listened to the audio samples again and rated the consonant feature accuracy. For general pronunciation accuracy, raters listened to

each word and evaluated the pronunciation of the word on a 9-point Likert scale from 9 = *definitely accurate* to 1 = *not accurate at all* by focusing on the production of both consonants and vowels as well as the lexical tones. For the accuracy of the consonantal feature, they were asked to pay attention to the word-initial segment and likewise evaluate it on a 9-point Likert scale. Raters were asked to focus on the strength of air burst for assessing the accuracy of aspiration features for plosives and on the duration of the airflow for assessing the accuracy of frication features for affricates. For instance, for /p^h/, raters should give a high score (7–9) if they heard a /p^h/-like sound with a strong air burst, a middle score (4–6) if they could not recognize whether the sound was /p/ or /p^h/, and a low score (1–3) for a /p/-like sound with a weak air burst. Additionally, raters assessed the accuracy of consonantal feature regardless of whether the consonant produced was the target consonant or not. For example, if the target sound was /t^ha/ but the participant produced /k^ha/, the rater still gave the output an "accurate" rating because the aspiration feature was accurately pronounced even if the consonant itself was not.

Interrater reliability of both accuracy measures was checked using Cronbach's alpha. The results revealed a good level of agreement for general pronunciation accuracy (α = .894) and an excellent level of agreement for consonantal feature accuracy (α = .928).

Musical background. As musical experience may play a role in phonological learning (see Chobert & Besson, 2013, for a review), we controlled for this factor to ensure that there was no difference in this regard between the two training groups. First, adapting Boll-Avetisyan et al.'s (2017) procedure, a musical expertise score was obtained for each participant by coding their answers to the musical background questionnaire as follows: (a) For the years spent studying music, the score equaled the number of years reported; (b) for the number of instruments played, 1 point was given for each instrument reported; and (c) regarding the amount of time they reported spending on a regular basis singing or listening to music, 5 points

were given if the participant reported 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." The sum of all these scores constituted a "musical expertise" score. In addition, the musical skill of participants was rated by means of a self-assessment tool (following Law & Zentner, 2012) that yielded a score whereby 1 = *nonmusician*, 2 = *music-loving non- musician*, 3 = *amateur musician*, 4 = *semiprofessional musician*, and 5 = *professional musician*. This constituted their "self-perceived musical skills" score.

Statistical Analyses

First, three independent-samples t tests were applied using IBM SPSS Statistics 25 software in order to check whether participants differed in terms of age, musical expertise, and self-assessed musical skill across the two between-subjects groups. The t-test results were as follows: (a) age, $t(48) = -0.729$, $p = .114$; (b) musical expertise, $t(48) = -0.048$, $p = .962$; and (c) self-perceived musical skills, $t(48) = -0.223$, $p = .825$. These results showed that the participants in the two between-subjects groups were not statistically different in terms of these three features.

A generalized linear mixed model (GLMM) was applied to the following outcome measures using the *glmmTMB* package (Brooks et al., 2017) in R: (a) identification accuracy, that is, the participant's score for each item in the identification task; (b) pronunciation accuracy, that is, the participant's general pronunciation score for each item rated by each rater in the imitation task; or (c) consonantal feature accuracy, that is, the participant's consonantal feature score for each item rated by each rater in the imitation task. The fixed factors were condition (two levels: NG and G), test (two levels: pre- and posttest), and consonant type (two levels: plosives and affricates), as well as their interactions. A series of GLMMs using different random effects structures were modeled, from the most complex random effects structure to a marginal model with no random effects. All the structures that did not produce any converge

problems were compared using the function *compare performance* from the *performance* package (Lüdecke et al., 2019) to identify the model that best fitted our data. For identification accuracy, the best-fitting model was the one including a random intercept both for participant and for item (i.e., (1 | Participant) + (1 | Item)). In the case of the pronunciation accuracy, the best-fitting model was the one with the random effects structure including random slopes for Test, Consonant Type and Test × Consonant Type by Participant, for Condition by Item, and for Consonant Type by Rater (i.e., (1 + Test * Consonant Type | Participant) + (1 + Condition | Item) + (1 + Consonant Type | Rater)). In the case of the consonantal feature accuracy, the best-fitting model was the one including a random slope for Test, Consonant Type and Test × Consonant Type by Participant, a random slope for both Condition and Test by Item, and a random intercept for rater (i.e., (1 + Test * Consonant Type | Participant) + (1 + Condition + Test | Item) + (1 | Rater)). In the results below, the omnibus test results are provided, plus the output of a series of Bonferroni pairwise tests performed with the *emmeans* package (Lenth et al., 2019), which include a measure of effect size by using Cohen's *d*.

Results

The random effect (co)variance of the three GLMMs is reported in Table 1. Furthermore, the results of the three GLMMs with the three outcome measures (identification accuracy, pronunciation accuracy, and consonantal feature accuracy) as the dependent variables are illustrated in Table 2.

Table 1. Random effects (co)variances of the three generalized linear mixed models on identification accuracy, pronunciation accuracy, and consonantal feature accuracy.

	Term	Name	*SD*
Identification Accuracy	ID	(Intercept)	0.116
(Number of observations: 1200)	Item	(Intercept)	0.068
	Residual		0.380

Pronunciation Accuracy	ID	(Intercept)	1.468
(Number of observations: 6000)		Test	1.097
		Consonant Type	0.840
		Test × Consonant Type	0.578
	Item	(Intercept)	0.405
		Condition	0.188
	Rater	(Intercept)	0.465
		Consonant Type	0.122
	Residual		1.233
Consonantal Feature Accuracy	ID	(Intercept)	3.390
(Number of observations: 6000)		Test	2.220
		Consonant Type	1.663
		Test × Consonant Type	1.167
	Item	(Intercept)	1.179
		Condition	0.565
		Test	0.637
	Rater	(Intercept)	0.348
	Residual		2.350

Table 2. Results of the three generalized linear mixed models on identification accuracy, pronunciation accuracy, and consonantal feature accuracy.

Term	Identification Accuracy		Pronunciation Accuracy		Consonantal Feature Accuracy	
	Chisq(df)	*Sig.*	*Chisq(df)*	*Sig.*	*Chisq(df)*	*Sig.*
Condition	1.117(1)	.291	0.713(1)	.399	0.624(1)	.430
Test	4.842(1)	.028*	0.019(1)	.889	0.526(1)	.468
Consonant Type	15.557(1)	< .001*	6.560(1)	.010*	5.271(1)	.022*
Condition × Test	0.697(1)	.404	0.262(1)	.609	0.001(1)	.989
Condition × Consonant Type	0.697(1)	.404	3.595(1)	.058	0.062(1)	.803

Test × Consonant Type	0.973(1)	.324	6.105(1)	.013*	2.244(1)	.134
Condition × Test × Consonant Type	2.078(1)	.149	5.854(1)	.016*	4.15(1)	.042*

Note. Significant results are marked with asterisk.

Effect of Hand Gestures on the Perception of Aspiration Contrasts

Figure 4 shows the mean identification accuracy rates across condition (NG and G), test (pretest and posttest), and consonant type (plosives and affricates). Results of the GLMM with the identification accuracy as the dependent variable (see Table 2) revealed a significant main effect of consonant type, $\chi^2(1) = 15.557$, $p < .001$, and test, $\chi^2(1) = 4.842$, $p = .028$, indicating that participants' performance differed significantly from pretest to posttest and varied significantly across the two types of consonant. No main effect of condition was found, $\chi^2(1) = 1.117$, $p = .291$. Post hoc comparisons revealed that participants improved significantly after training in general ($d = 0.13$, $p = .028$) and identified plosives significantly better than affricates ($d = 0.47$, $p < .001$). No significant two-way interaction between Condition × Test, $\chi^2(1) = 0.697$, $p = .404$, or the three-way interaction between Condition × Test × Consonant Type, $\chi^2(1) = 2.078$, $p = .149$, was found, which suggests that (a) training with gestures did not yield more of an improvement for the perception of nonnative aspiration contrast than training without gestures and (b) the appropriateness of the gesture (i.e., whether or not the form or manner of the hand gesture seemed to visually mimic the phonetic properties of the target phonemes) did not influence the training effect.

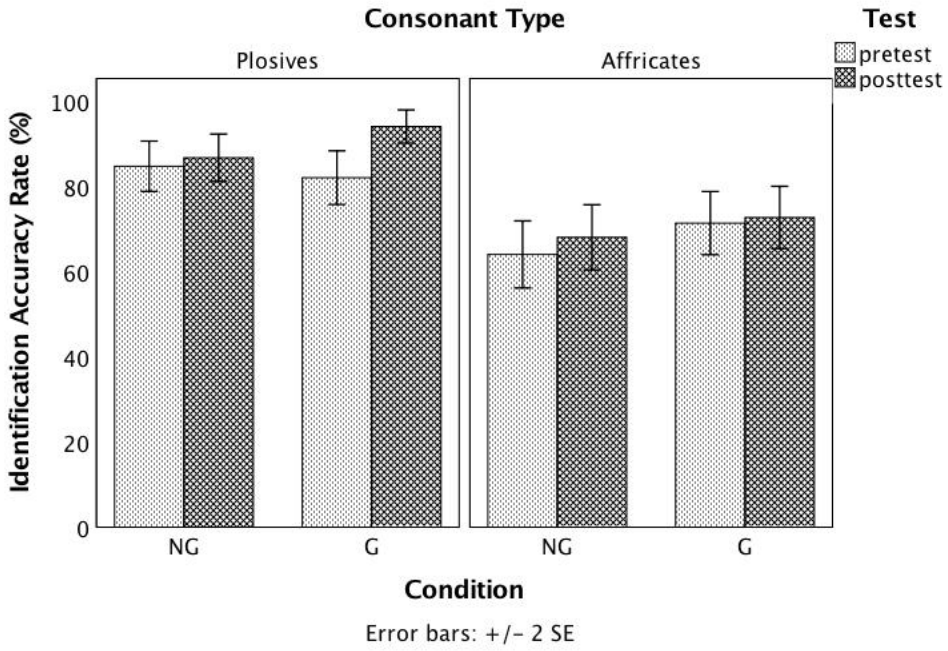


Figure 4. Mean identification accuracy rates across condition (No Gesture [NG] and Gesture [G]), test (pretest and posttest), and consonant type (plosives and affricates). Error bars indicate 2 *SE*s.

Effects of Hand Gestures on the Production of the Aspiration Contrasts

Pronunciation accuracy. Figure 5 shows the mean pronunciation accuracy obtained in the imitation task across condition (NG and G), test (pretest and posttest), and consonant type (plosives and affricates). Results of the GLMM with pronunciation accuracy as the dependent variable (see Table 2) revealed a main effect of consonant type, $\chi^2(1) = 6.560$, $p = .010$, showing that participants' general pronunciation was significantly different depending on whether they pronounced plosives or affricates. Participants showed better performance on the plosives than on the affricates ($d = 0.391$, $p = .007$). A significant two-way interaction was found, that is, Test × Consonant Type, $\chi^2(1) = 6.105$, $p = .013$, indicating that participants' performance on different types of consonants differed significantly from pre- to posttest. Crucially, a significant three-way interaction was found in Condition × Test × Consonant Type, $\chi^2(1) = 5.854$, $p = .016$, revealing that the participants in both the G and NG groups were significantly different from

pretest to posttest depending on the type of consonant. Post hoc comparisons (see Table 3) showed that participants' general pronunciation of the words containing plosives improved significantly in the G group (d = 0.270, p = .023) but not in the NG group, while for words containing affricates, performance did not improve in either the NG or G group. These results showed that training with gestures that appropriately mimicked phonetic features significantly helped the general pronunciation of the target words compared to training without gestures and training with gestures that inappropriately bore phonetic behaviors. These results suggest that, when gestures appropriately represent the auditory and articulatory properties of the target phonemes, participants can improve their pronunciation performance. By contrast, if a gesture does not visually match the auditory and articulatory properties of the target phonemes, this may not help the pronunciation.

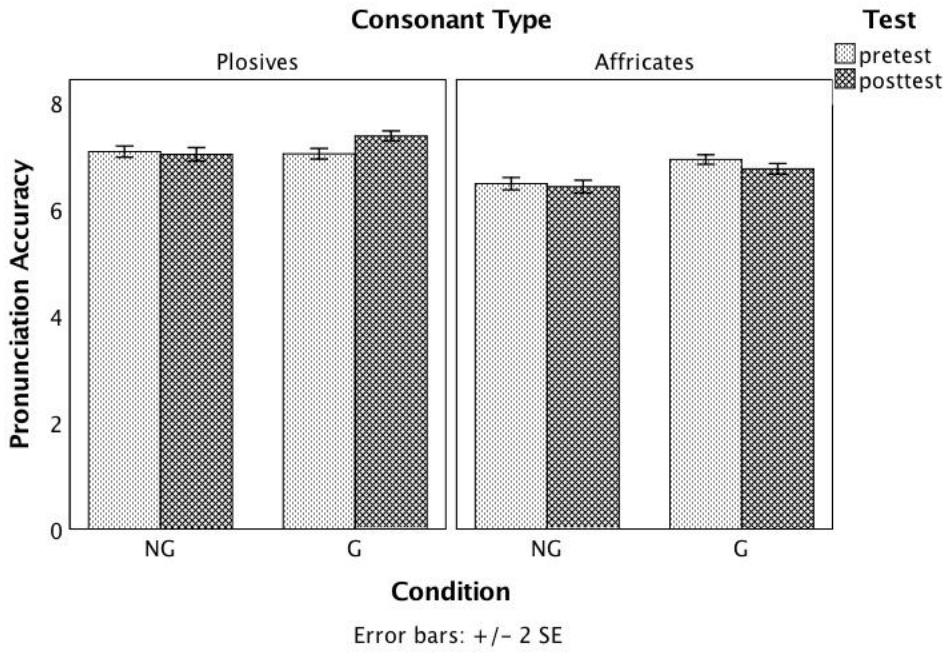


Figure 5. Mean pronunciation accuracy across condition (No Gesture [NG] and Gesture [G]), test (pretest and posttest), and consonant type (plosives and affricates). Error bars indicate 2 *SE*s.

Table 3. Post hoc results of the pairwise contrast at pretest and posttest according to

condition and consonant type.

		Pronunciation Accuracy		*Consonantal Feature Accuracy*	
		Coh. d	*Sig.*	*Coh. d*	*Sig.*
Plosives	G	0.270	.023*	0.339	.032
	NG	-0.041	.729	0.092	.560
Affricates	G	-0.141	.186	-0.107	.434
	NG	-0.045	.669	0.001	.991

Note. Significant results are marked with asterisk.

Consonantal feature accuracy. Figure 6 shows mean consonantal feature accuracy across condition (NG and G), test (pretest and posttest), and consonant type (plosives and affricates). Results of a GLMM with consonantal feature accuracy as the dependent variable (see Table 2) revealed a main effect of consonant type, $\chi^2(1) = 5.271$, $p = .022$, suggesting that participants' pronunciation performance of the two types of consonants differed significantly. A significant three-way interaction between Condition × Test × Consonant Type, $\chi^2(1) = 4.150$, $p = .042$, was obtained. The post hoc results (see more details in Table 3) showed that the improvement from pre- to posttest for plosives was present only in the G condition ($d = 0.339$, $p = .032$), but not for any combination including NG or affricate data, suggesting that participants' production of consonantal features improved only when gestures matched the properties of the target phonemes. When participants were trained with matching gestures, their performance on consonantal feature improved significantly from pretest to posttest; however, when they were trained with mismatching gestures or without any gestures, their performance did not improve after training. Taken together, these results strongly suggest that training with

gestures had asymmetric effects depending on whether the gestures appropriately mimicked the phonetic features of the target phonemes or not.

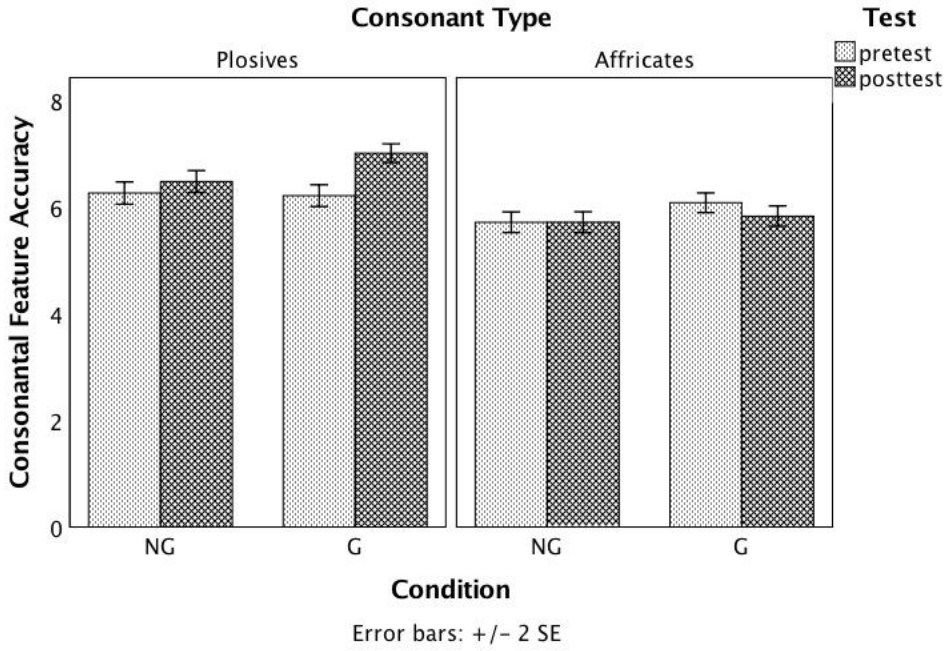


Figure 6. Mean consonantal feature accuracy across condition (No Gesture [NG] and Gesture [G]), test (pretest and posttest), and consonant type (plosives and affricates). Error bars indicate 2 SEs.

Discussion and Conclusions

This study examined whether observing hand gestures mimicking the aspiration feature of aspirated consonants can facilitate the initial-stage learning of nonnative consonants from two complementary dimensions, namely, the *perception* of aspirated and unaspirated consonants (which was examined through an identification task) and the *production* of the aspiration contrasts (which was tested through an imitation task). While previous studies have shown beneficial effects of pitch gestures on the learning of suprasegmental features, such as lexical tones (e.g., Baills et al., 2019) and intonation patterns (Yuan et al., 2019), and also beneficial effects of beat gestures on rhythmic patterns (Gluhareva & Prieto, 2017), more inconsistent results have been found for the role of phonetically based gestures on the learning

of L2 segmental phonology (Amand & Touhami, 2016; Hoetjes et al., 2019). Crucially, this study examined whether the appropriateness of the gesture relative to the phonetic features of the aspirated consonants would impact, in any way, the training effect. In our experimental setting, while a hand burst gesture visually mimicked the perceptual auditory and articulatory features of aspirated plosives, it did not match those of aspirated affricates given their more prolonged frication phase and lack of a prominent air burst. Thus, while the hand burst gesture used in this study visually matched the burst of air related to the aspiration feature in the case of aspirated plosives, this was not the case for aspirated affricates.

Results of the identification task showed that both training groups (G and NG) made a significant improvement in their perception of nonnative phonemic contrasts from pre- test to posttest. However, training with gestures did not yield a more significant improvement in identification accuracy than training without gestures. Importantly, training with hand burst gestures had a similar effect on the two types of consonants. Thus, on the one hand, these results suggest that the appropriateness of gestures in terms of their visual match to the phonetic properties of the target phonemes may not influence the training effect on perception.

By contrast, the results obtained for pronunciation accuracy in the imitation task showed that improvement between pre- and posttest crucially depended on whether the gesture appropriately mimicked the consonantal feature. While the use of the hand burst gesture in the G condition was effective for the pronunciation of plosives in terms of both aspiration feature accuracy and general pronunciation accuracy, no beneficial effects were found in the NG condition. Specifically, no gains were found for accuracy in the production of the consonantal feature and the general pronunciation of the affricates. Considering that the hand burst gesture matched the aspiration feature of the plosive only, these results strongly suggest that gestures may enhance the learning of nonnative segments only when they appropriately depict the phonetic information of the target segments.

Moreover, these results allow us to reconcile recent contradictory results reported on the effects of gestural training on target segmental features. While some gestures mimicking the phonetic feature of target phonemes have been shown to facilitate nonnative speakers' production of released and unreleased stops in English (Amand & Touhami, 2016), as well as the Spanish phoneme /u/ (Hoetjes et al., 2019), this was not the case when the technique was applied to Spanish /θ/ (Hoetjes et al., 2019). In our view, while the gestures used for English stops and the Spanish phoneme /u/ provide clearly visible articulatory information about the target sound, this is not the case for the gesture used for the phoneme /θ/, which depicts the tongue shape during the production of this consonant but lacks crucial information about the location of the tongue relative to the teeth or jaw aperture. Thus, hand gestures may not have a positive effect on learning if they depict relatively unimportant phonetic information (e.g., the flat hand gesture for /θ/ in Hoetjes et al., 2019) or are incongruent with the phonetic feature in question (e.g., the burst gesture for affricates, as in our study). A similar observation was made by Hirata and Kelly (2010), who pointed out that the specific choice of hand gestures might have been the reason for the negative effect of gestures on the learning of nonnative durational vowel contrasts in Japanese. In their study, beat gestures were used to teach short vowels and sweep gestures were used to teach long vowels. However, the use of these gestures did not yield a significant improvement in the perception of these vowels in comparison with not providing gestures. While beat gestures showed benefits for the learning of nonnative prominence marking (Gluhareva & Prieto, 2017; Kushch, 2018; Llanes-Coromina et al., 2018), this type of gesture may not be appropriate to represent a short vowel that is nonprominent. While previous studies have shown that hand gestures could facilitate the learning of L2 word and lexical tones only when the gesture form matched the learning targets (e.g., Kelly et al., 2009; Zhen et al., 2019), our study has further explored the influence of the correspondence between gesture form and the phonetic features of L2 speech sounds on the learning of L2

segments. Our study suggested that the success of certain gesture types for the acquisition of segmental information may depend on how well the gesture serves as a visual representation of visible and audible target speech information.

In this study, contrary to expectation, while training with gestures clearly led to pronunciation gains, it had limited effects on perception. There are several possible explanations for this result. First, the hand gesture may have asymmetric effects on perception and production in phonological processing. For perception, participants were tested on how well they were able to hear subtle differences in the degree of aspiration. During training, participants in the G group may have paid more attention to the very prominent visual signals provided by the hand gestures than to the auditory distinctions between the aspiration contrasts. Participants in the NG group, on the other hand, may have been more attuned to the auditory channel, since the visual distinction from lip movements of the aspiration contrasts is subtle. Second, our training session lasted only 5 min 36 s, a relatively short time if compared to other perceptual training studies (e.g., 400 min in Hazan et al., 2005; 120 min in Hirata & Kelly, 2010). Thus, a longer training period could perhaps have led to clearer effects on perception and might have facilitated further the building of new phonological categories by naive learners. Third, the current study showed that the production abilities of participants in the G group did not improve linearly with their perception ability, which is consistent with the results of previous phonetic training studies on English /l/–/r/ (e.g., Hazan et al., 2005) or French nasal vowels (e.g., Inceoglu, 2016). The relationship between L2 perception and production development has been studied for decades, yet it is still not clear. While some studies suggest that, during perceptual training, perception precedes production, others support that production precedes the development of perception (see Sakai & Moorman, 2018, for a review).

Several limitations and future research questions related to this study can be identified. First, even though our training focused on specific consonantal features, we found that

participants significantly improved in terms of not only the accuracy of those particular features but also pronunciation in general. Still, it remains to be seen whether this general improvement in pronunciation was due exclusively to a greater skill with the consonantal features of the target words or also reflected improvement in the handling of other phonetic features of the target words, in other words, whether the segment-focused training led to collateral improvement in nontarget phonological features. Second, although participants in the G group improved significantly in terms of both aspiration and general pronunciation after training, this training focused on perception and did not include imitation practice. Future studies could assess whether training sessions in which participants are asked to imitate gestures while repeating words lead to greater improvement. Third, although participants in the G group improved their production of plosives immediately after training, it remains to be tested whether the training could encourage retention of learning effects over time. Thus, a future study could assess the retention of training with hand gestures with a delayed posttest. In addition, since a beneficial effect of hand burst gestures on the learning of Mandarin plosives has been detected in this study within a controlled laboratory setting, it would be of interest to examine the possible supporting role of gestures on other types of nonnative phonemes and also whether similar effects can be obtained in classroom settings. Finally, although hand gestures show beneficial effects on learning different sorts of speech sounds and they provide additional visual input of invisible articulatory features of certain phonemes (e.g., the aspiration feature, the tongue shape), it is important to mention that not all phonetic properties of phonemes could be represented by hand gestures (e.g., the vocal vibration feature for phonemes contrasting in voicing).

To conclude, this study shows that, for naive L2 learners, observing gestures that appropriately mimic phonetic processes can be effective to help learners produce nonnative segmental contrasts. From a pedagogical perspective, our findings have implications for L2

pronunciation teaching methods, as they clearly suggest that the incorporation of a multimodal approach to teaching L2 pronunciation patterns at the segmental level would be beneficial. In conjunction with previous research (e.g., Bailly et al., 2019), the results of this study suggest that the use of gestures that appropriately represent phonetic features of the target language can constitute a helpful resource for facilitating the pronunciation learning processes at early stages of phonological encoding. All in all, since this work has also revealed that the efficacy of specific gesture forms is constrained by the degree of match between the form of the hand gestures and the cued phonetic features, more work will be needed to assess the value of manual movements in relation to specific target phonetic features.

Acknowledgments

This research was supported by funding from the Spanish Ministry of Science, Innovation and Universities (PGC2018- 097007-B-I00) and the Generalitat de Catalunya projects (2017 SGR-971). The third author holds a predoctoral research grant awarded by the Department of Translation and Language Sciences, Universitat Pompeu Fabra.

The authors sincerely thank Jun Tang, Jinsong Wang, Songlin Wang, Xuejie Wang, and Siyu Zhou for their participation in the ratings. Many thanks also to Joan C. Mora (University of Barcelona) and Mario Bisiada (Universitat Pompeu Fabra) for their comments and suggestions on the first draft of this article and to their statistician, Joan-Borràs Comes (Universitat Pompeu Fabra - Autonomous University of Barcelona), for his crucial help on data analysis.

References

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

Arnold, P., & Hill, F. (2001). Bisensory augmentation: A speechreading advantage when speech is clearly audible and intact. *Journal of Psychology, 92*(2), 339-355. https://doi.org/10.1348/000712601162220

Baills, F., Suárez-González, N., González-Fuente, S., & Prieto, P. (2019). Observing and producing pitch gestures facilitates the learning of Mandarin Chinese tones and words. *Studies in Second Language Acquisition*, *41*(1), 33-58. https://doi.org/10.1017/S0272263118000074

Boll-Avetisyan, N., Bhatara, A., & Höhle, B. (2017). Effects of musicality on the perception of rhythmic structure in speech. *Laboratory Phonology*, 8, 1–16. http://doi.org/10.5334/labphon.91

Brooks ME, Kristensen K, van Benthem KJ, Magnusson A, Berg CW, Nielsen A, Skaug HJ, Maechler M, & Bolker BM. (2017). glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling. *The R Journal, 9*(2), 378-400. Retrived from https://journal.r-project.org/archive/2017/RJ-2017-066/index.html

Chen, N. F., Shivakumar, V., Harikumar, M., Ma, B., & Li, H. (2013). Large-scale characterization of Mandarin pronunciation errors made by native speakers of European languages. Paper presented at the 14th Annual Conference of the International Speech Communication Association, Lyon, France. Retrived from https://www.isca-speech.org/archive/archive_papers/interspeech_2013/i13_2370.pdf

Chobert, J., & Besson, M. (2013). Musical expertise and second language learning. *Brain Sciences, 3*(2), 923-940. https://doi.org/10.3390/brainsci3020923

Duanmu, S. (2007). *The phonology of standard Chinese*. New York, NY: Oxford University Press.

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

Grant, K.W., & Seitz, P.F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of Acoustical Society of America, 108*(3), 1197-1208. https://doi.org/10.1121/1.1288668

Grant, K.W., Walden, B.E., & Seitz, P.F. (1998). Auditory-visual speech recognition by hearing-impaired subjects: consonant recognition, sentence recognition, and auditory-visual integration. *The Journal of Acoustical Society of America, 103*(5), 2677-2690. https://doi.org/10.1121/1.422788

Hannah, B., Wang, Y., Jongman, A., Sereno, J. A., Cao, J., & Nie, Y. (2017). Cross-modal association between auditory and visuospatial information in Mandarin tone perception in noise by native and non-native perceivers. *Frontiers in psychology, 8*:2051. https://doi.org/10.3389/fpsyg.2017.02051

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. W., van Maastricht, L. J., & Heijden, L. (2019). Gestural training benefits L2 phoneme acquisition: findings from a production and perception perspective. *Proceedings of the 6th Gesture and Speech in Interaction Conference*, Paderborn, Germany. Retrieved from https://hdl.handle.net/2066/209532

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

Kelly, S. D., Bailey, A., & Hirata, Y. (2017). Metaphoric gestures facilitate perception of intonation more than length in auditory judgments of non-native phonemic contrasts. *Collabra: Psychology, 3*(1): 7. http://doi.org/10.1525/collabra.76

Kelly, S.D., & Hirata, Y. (2017). What neural measures reveal about foreign language learning of Japanese vowel length contrasts with hand gestures. In S. Tanaka et al. (Eds.), *音韻研究の新展開：窪薗晴夫教授還暦記念論文集 [New Development in Phonology Research: Festschrift in Honor of Haruo Kubozono]*. Tokyo, Japan: Kaitakusha.

Kelly, S. D., McDevitt, T., & Esch, M. (2009). Brief training with co-speech gesture lends a hand to word learning in a foreign language. Language and cognitive processes, 24(2), 313-334. https://doi.org/10.1080/01690960802365567

Klick, S. L. (1985). Adapted cuing technique for use in treatment of dyspraxia. *Language, Speech, and Hearing Services in Schools, 16*(4), 256-259. https://doi.org/10.1044/0161-1461.1604.256

Kushch, O. (2018). *Beat gestures and prosodic prominence: impact on learning* (Doctoral dissertation, Universitat Pompeu Fabra, Barcelona, Spain). Retrieved from http://hdl.handle.net/10803/463004

Ladefoged, P. (1963). Some physiological parameters in speech. *Language and speech, 6*(3), 109-119. https://doi.org/10.1177/002383096300600301

Laver, J. (1994). *Principles of Phonetics*. New York, NY: Cambridge University Press.

Law, L. N., & Zentner, M. (2012). Assessing musical abilities objectively: Construction and validation of the Profile of Music Perception Skills. *PloS one*, *7*(12), e52508.

Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2019). Emmeans: Estimated marginal means, aka least-squares means. R package. See https://CRAN.R-project.org/package=emmeans

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,* 1-25. https://doi.org/10.1017/S0272263120000054

Llanes-Coromina, J., Prieto, P., & Rohrer, P. L. (2018). Brief training with rhythmic beat gestures helps L2 pronunciation in a reading aloud task. *Proceedings from the 9th International Conference on Speech Prosody*, Poznań, Poland. http://dx.doi.org/10.21437/SpeechProsody.2018-101

Lüdecke, D., Makowski, D., Waggoner, P., & Patil, I. (2019). Performance: Assessment of regression models performance. R package. See https://CRAN.R-project.org/package=performance

Macedonia, M., Müller, K., & Friederici, A. D. (2011). The impact of iconic gestures on foreign language word learning and its neural substrate. *Human brain mapping*, 32(6), 982-998. https://doi.org/10.1002/hbm.21084

Morett, L. M., & Chang, L.Y. (2015). Emphasising sound and meaning: Pitch gestures enhance Mandarin lexical tone acquisition. *Language, Cognition and Neuroscience, 30*, 347-353. https://doi.org/10.1080/23273798.2014.923105

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746-748. https://doi.org/10.1038/264746a0

Navarra, J., & Soto-Faraco, S. (2007). Hearing lips in a second language: Visual articulatory information enables the perception of second language sounds. *Psychological Research, 71*(1), 4-12. http://doi.org/10.1007/s00426-005-0031-5

Rusiewicz, H. L., & Rivera, J. L. (2017). The effect of hand gesture cures 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

Shelton, I. S., & Garves, M. M. (1985). Use of visual techniques in therapy for developmental apraxia of speech. Language, Speech, and Hearing Services in Schools, 16(2), 129-131. https://doi.org/10.1044/0161-1461.1602.129

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

Summerfield, Q. (1992). Lipreading and audio-visual speech perception. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 335*(1273), 71-78. https://doi.org/10.1098/rstb.1992.0009

Wheeler, M.W. (2005). *The phonology of Catalan.* New York, NY: Oxford University Press.

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

Zhang, Y. (2002). The importance of using gestures in pronunciation teaching. *Language Teaching and Linguistics Studies, 6*, 51-56.

Zhen, A., Van Hedger, S., Heald, S., Goldin-Meadow, S., & Tian, X. (2019). Manual directional gestures facilitate cross-modal perceptual learning. *Cognition*, *187* (2019), 178-187. https://doi.org/10.1016/j.cognition.2019.03.004

Appendix

Linguistic and Musical Background Questionnaire (English Translation)

Linguistic experience

1. What percentage of CATALAN do you use in your daily life?

2. Apart from CATALAN and SPANISH, which what language(s) do you speak?

3. Have you ever studied Japanese?

Musical experience

1. How many years of musical education have you ever received?

2. Do you play any instruments? If your answer is yes, answer. If not, move on to 4.

3. Which instrument(s) do you play?

4. 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

5. Which one of the following best describes you?

    A. I'm a non-musician

    B. I'm a music-loving non-musician

    C. I'm an amateur musician

    D. I'm a semi-professional musician

    E. I'm a professional musician

Supplemental materials

S1. Twelve Chinese disyllabic words used for the training session

Word	Target	Consonant Type	Catalan adaptation	IPA	Meaning
***p**í yán*	p^h	plosive	*p^hi yan*	p^hi jan	dermatitis
***b**í yán*	p	plosive	*pi yan*	pi jan	rhinitis
***t**ān liàn*	t^h	plosive	*t^han lian*	t^han lian	greed
***d**ān liàn*	t	plosive	*tan lian*	tan lian	one-side love
***k**ǒu liáng*	k^h	plosive	*k^hou liang*	k^hou liaŋ	ration
***g**ǒu liáng*	k	plosive	*kou liang*	kou liaŋ	dog food
***c**uò wù*	ts^h	affricate	*ts^huo u*	ts^huo u	error
***z**uò wù*	ts	affricate	*tsuo u*	tsuo u	crop
***q**īng lǐ*	$tɕ^h$	affricate	*tsj^{h}ing li*	$tɕ^h$iŋ li	to clean
***j**īng lǐ*	tɕ	affricate	*tsjing li*	tɕiŋ li	manager
***ch**ù lì*	$tʂ^h$	affricate	*tx^{h}u li*	$tʂ^h$u li	to stand
***zh**ù lì*	tʂ	affricate	*txu li*	tʂu li	boosting

Note. In the "Word" column, all items are written in *pinyin* (standard Romanized Chinese).
The IPA were transcribed based on Xiandai Hanyu [Modern Chinese] which is edited by Huang, B. and Liao, X. (2002).

S2. Twelve Chinese disyllabic words used for the identification task

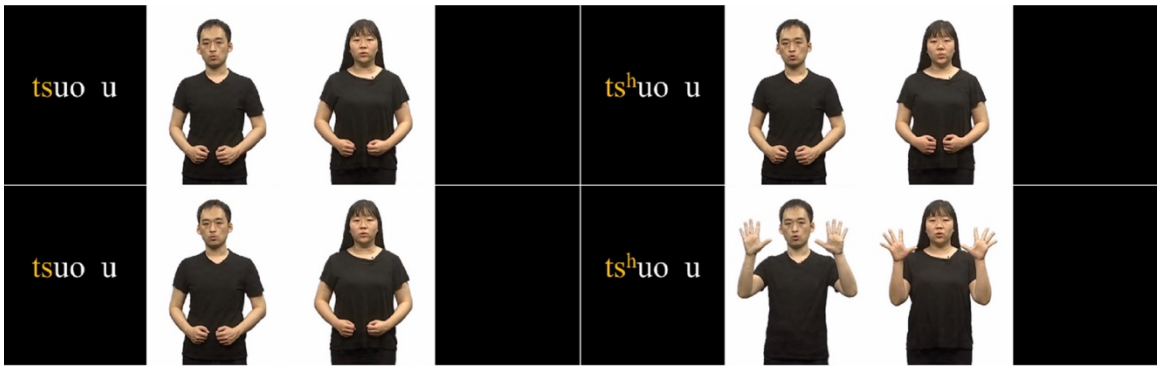
Word	Target	Trained	Catalan adaptation	IPA	Meaning
***p**í yán*	p^h	Yes	*p^hi yan*	p^hi jan	dermatitis
***b**í yán*	p	Yes	*pi yan*	pi jan	rhinitis
***t**ān liàn*	t^h	Yes	*t^han lian*	t^han lian	greed
***d**ān liàn*	t	Yes	*tan lian*	tan lian	one-side love
***k**ǔ lì*	k^h	No	*k^hu li*	k^hu li	labor
***g**ǔ lì*	k	No	*ku li*	ku li	encouragement
***c**uò wù*	ts^h	Yes	*ts^huo u*	ts^huo u	error

zuò wù	ts	Yes	*tsuo u*	tsuo u	crop
***q**ié hé*	tɕʰ	No	*tsjʰe he*	tɕʰiɛ xɤ	stuffed eggplant
***j**ié hé*	tɕ	No	*tsje he*	tɕiɛ xɤ	tuberculosis
***ch**ǎn shì*	ʈʂʰ	No	*txʰan xi*	ʈʂʰan ʂɻ	to elucidate
***zh**ǎn shì*	ʈʂ	No	*txan xi*	ʈʂan ʂɻ	to demonstrate

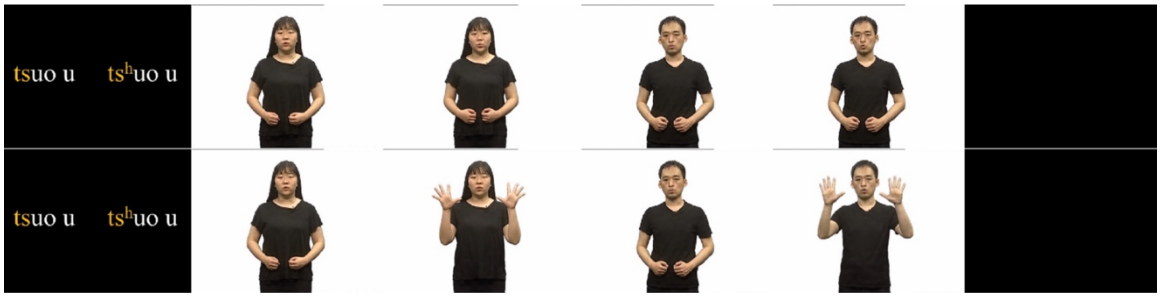
S3. Twelve Chinese disyllabic words used for the imitation task

Word	**Target**	**Trained**	**IPA**	**Meaning**	**Words**	**Target**	**Trained**	**IPA**	**Meaning**
***p**á shǒu*	pʰ	No	pʰa xou	thief	***c**ōng yóu*	tsʰ	No	tsʰuŋ jou	scallion oil
***b**á shǒu*	p	No	pa xou	handle	*z**ō**ng yóu*	ts	No	tsuŋ jou	palm oil
***t**ào lù*	tʰ	No	tʰau lu	strategy	***q**īng lǐ*	tɕʰ	Yes	tɕʰiŋ li	to clean
***d**ào lù*	t	No	tau lu	road	***j**īng lǐ*	tɕ	Yes	tɕiŋ li	manager
***k**ǒu liáng*	kʰ	Yes	kʰou liaŋ	ration	***ch**ù lì*	ʈʂʰ	Yes	ʈʂʰu li	to stand
***g**ǒu liáng*	k	Yes	kou liaŋ	dog food	***zh**ù lì*	ʈʂ	Yes	ʈʂu li	boosting

S4. Screenshots of one trial of the first training block in the NG condition (upper panel) and G condition (lower panel). Image used with consent from the instructors.



S5. Screenshots of one trial of the second training block in NG condition (upper panel) and G condition (lower panel). Image used with consent from the instructors.


tsuo u
tsʰuo u
tsuo u
tsʰuo u