Embodied Pronunciation Training:

The benefits of visuospatial hand gestures

Peng Li

TESI DOCTORAL UPF / 2021

DIRECTOR DE LA TESI

Dra. Pilar Prieto

DEPARTAMENT DE TRADUCCIÓ I CIÈNCIES DEL

LLENGUATGE

To my parents

To my grandmother

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my supervisor, Dr. Pilar Prieto, for her continuous support and patience. This PhD dissertation would have been impossible without her motivation and immense knowledge. I was keen to pursue a career as a linguistic researcher, which drove me from my homeland to Barcelona. During this beautiful journey, Dr. Prieto has been here, compassing me in exploring the linguistic world. The most valuable belief that I learned from her is that one always deserves to be a better self than she/he believed to be.

I am also grateful to Prof. Beijie Xia, the former head of the Department of Foreign Languages of Hefei University, where I worked four years ago. She made me a qualified language teacher and motivated me to pursue second language research. I sincerely appreciate the seminars she organized, which opened the door of SLA to me, and her friendly encouragement, support, and kind recommendations when I decided to start a PhD study.

I owe a sincere thanks to Dr. Lorraine Baqué (Universitat Autònoma de Barcelona), who has collaborated with us in the French study. Her great knowledge in French phonology, verbotonal method, and second language acquisition and her teaching experience has given me so much inspiration.

I would also like to thank the experts in the academic community who have given me great suggestions and insightful comments during my PhD study: Dr. Joan C. Mora, Dr. Kazuya Saito, Dr. Núria Esteve-Gibert, among many others.

Many thanks are devoted to the staff and students at the Department of Translation and Language Sciences. Thank you, Dr. Gemma Barberà, for being my tutor. Thank you, Dr. Joan Costa, Dr. Elisabeth Miche, and Ms. Vanessa Alonso, for helping me recruit participants. Thank you, all the participants, for their cooperation and dedication. Thank you, the technicians at the audiovisual studio of UPF, and the librarians for helping me record the audiovisual materials. This dissertation would never be realized without your help. Finally, a special thanks go to Mr. Rafa Ordóñez and Ms. Núria Abad for their help in the administrative affairs.

I thank all my friends who voluntarily rated the speech data without asking for any compensation: Ruochen Ning, Luo Wang, Songlin Wang, Yuan Zhang, and Siyu Zhou. I also thank my dearest colleagues Patrick L. Rohrer, Ïo Valls, and Ingrid Vilà, as well as our friend Pedro Navarro for the working memory materials.

To my lovely colleagues and co-authors, Florence Baills and Xiaotong Xi, as well as the rest of the GrEP members, Júlia Florit, Mariia Pronina, Patrick L. Rohrer, Ïo Valls, Ingrid Vilà, and Yuan Zhang, it is such a

pleasure to work with you. I will always cherish the coffee and lunch time we spent together, the brainstorms during our seminars, and the kindest help you offered whenever I needed it.

Huge thanks to all my beloved friends: Qiang Feng, Meishen Liu, Meztli Santamaría, Chengjia Ye, and Siyi Zhao, for all the laughter and joy we share. During my PhD study, whenever I had a hard time, you would always cheer me up. Thank you, Mengran An, Zhongrui Bian, Yi Cheng, Anbang Jing, Taikun Li, Yajie Li, Wei Song, Qingyuan Sun, Weibo Xu, Tingfang Yu, Xuezheng Yin, Cancan Yuan, and Jianfeng Zhao, for all the best memories and your friendship. It has been the energy for the current me and will always be the power for the future me.

I must thank my parents for their support of my academic career. I know it has been hard for them without me being at home, but they never complained about my absence. They wanted me to pursue my ideal life so that they unconditionally supported me with every decision I made, even if it meant that I would be away from home. Now the big end is coming, and I will be back home soon! I have to thank my dear Siyu, for being with me during these years. I cannot imagine how would life in Barcelona be without you here. This life-long memory will always be shimmering in my heart, for now, and forever.

Table of Contents

ABSTRACT

In the last few decades, the use of hand gestures that encode phonological features of the target language has been proven to play a positive role in L2 suprasegmental learning. However, less is known about the effects of embodied pronunciation training on the acquisition of novel segments. This doctoral dissertation includes three between-subject studies which tested the effects of visuospatial hand movements as pedagogical gestures for training L2 pronunciation features.

Study 1 demonstrated that producing durational gestures (i.e., horizontal hand movements to illustrate vowel-length contrasts) improves novice learners' production of Japanese long vowels. Study 2 showed that appropriately performing gestures that mimic consonantal aspiration boosts the learning of Mandarin aspirated plosives by novice learners. Finally, study 3 revealed that the observation of hand gestures encoding melodic and rhythmic features of speech helps learners with elementary-to-intermediate French proficiency reduce their accentedness and improve their accuracy in producing the non-native front rounded vowels.

Overall, the three studies show the benefits of embodied pronunciation training involving hand gestures that encode segmental and suprasegmental phonological information. These results highlight the need to in-

tegrate embodied training methods in L2 classrooms and support the predictions of the Embodied Cognition paradigm for L2 phonological learning.

Key words: Gestures, hand movements, pronunciation, embodied cognition, second language acquisition

RESUM

Al llarg de les últimes dècades s'ha demostrat que emprar gestos manuals que fan visibles els aspectes fonològics d'una llengua estrangera afavoreix l'aprenentatge de la prosòdia d'aquesta llengua. No obstant això, hi ha menys estudis sobre l'efectivitat d'aquests gestos en l'adquisició de nous sons. Aquesta tesi doctoral inclou tres estudis experimentals que tenen per objectiu avaluar els efectes dels entrenaments multimodals de la pronúncia que inclouen aquests moviments manuals com a gestos pedagògics per a l'entrenament dels trets fonètics .

L'estudi 1 demostra que emprar gestos que codifiquen trets fonètics de duració dels sons (i.e., moviments horitzontals de les mans que il·lustren aquests contrastos de durada vocàlica) millora la producció de les vocals llargues del japonès per part d'aprenents novells. L'estudi 2 mostra que una realització adequada dels gestos que imiten els trets d'aspiració consonàntica facilita l'aprenentatge de les consonants oclusives aspirades del xinès per part de nous aprenents. L'estudi 3 demostra que un entrenament multimodal que integri gestos manuals que facin visibles els trets prosòdics del francès ajuda els aprenents d'aquesta llengua a reduir el seu accent i alhora augmentar la precisió en la pronúncia de les vocals arrodonides anteriors.

En resum, els tres estudis mostren els beneficis de la pràctica multimodal de la pronúncia amb exercicis que incloguin gestos manuals que codifiquen informació fonològica a nivell segmental i suprasegmental. Els resultats ressalten la importància d'incorporar entrenaments multimodals de la pronúncia en l'aula de llengües estrangeres i donen suport a les prediccions del paradigma de la Cognició Corporeïtzada (*Embodied Cognition*) sobre l'aprenentatge fonològic de segones llengües.

Paraules clau: Gestos, moviments manuals, pronúncia, Cognició Corporeïtzada, adquisició de segones llengües

RESUMEN

A lo largo de las últimas décadas, se ha demostrado que el uso de gestos manuales que visualizan aspectos fonológicos de una segunda lengua facilita el aprendizaje de la prosodia de esta lengua. No obstante, hay menos estudios sobre la efectividad de esos gestos en la adquisición de nuevos sonidos. Esta tesis doctoral incluye tres estudios experimentales que tienen por objetivo evaluar los efectos de entrenamientos multimodales de la pronunciación que incluyen esos movimientos manuales como gestos pedagógicos de los rasgos fonéticos.

El estudio 1 demuestra que el uso de gestos que codifican rasgos fonéticos de duración de los sonidos (i.e., movimientos horizontales de las manos que ilustran los contrastes de duración vocálica) mejora la producción de las vocales largas del japonés por parte de nuevos aprendices. El estudio 2 muestra que una realización adecuada de los gestos que imitan los rasgos de aspiración consonántica facilita el aprendizaje de las consonantes oclusivas aspiradas del chino por parte de estudiantes principiantes. El estudio 3 demuestra que un entrenamiento multimodal que incluye el uso de gestos manuales que codifican los rasgos prosódicos del francés ayuda a los estudiantes de esta lengua a reducir su acento y aumentar la precisión en la pronunciación de las vocales labializadas anteriores.

En resumen, los tres estudios muestran los beneficios de la práctica multimodal de la pronunciación con ejercicios que incluyan gestos manuales que codifican información fonológica a nivel segmental y suprasegmental. Los resultados resaltan la importancia de incorporar entrenamientos multimodales en el aula de lenguas extranjeras y apoyan las predicciones del paradigma de la Cognición Corporeizada (*Embodied Cognition*) en el contexto del aprendizaje fonológico de segundas lenguas.

Palabras clave: Gestos, movimientos manuales, pronunciación, Cognición Corporeizada, adquisición de segundas lenguas

1

CHAPTER 1:
GENERAL INTRODUCTION

1.1 Theoretical framework

1.1.1 The Embodied Cognition paradigm

In the past two decades, several theoretical frameworks within the cognitive sciences have rejected the traditional view that the mind and the body are two separated systems and have tried to capture the strong relation between mind and body. The Embodied Cognition paradigm (Foglia & Wilson, 2013; Ionescu & Vasc, 2014) holds that cognitive processes do not just compute amodal symbols in a modular system but are instead grounded in the sensory-motor processes and the internal states of the human body. Empirical evidence has shown how body movements can help people better encode and retain the information being perceived. For example, it has been observed that body movements may facilitate the conceptualization of abstract meaning (Barsalou, 2008, 2010) and boost the development of cognitive functions (Borghi & Caruana, 2015) and that embodied experiences can aid information recall (Kontra et al., 2015; Mizelle & Wheaton, 2010). In this sense, the body is viewed as an extension of our mind.

The Embodied Cognition paradigm emphasizes the role that "action" plays in cognitive processes in the sense that they are strongly dependent on sensory-motor experience. Further, a close link between action and language processing has been empirically demonstrated, particularly in relation to lexical processing (Pulvermüller et al., 2005), lexical recognition (Myung et al., 2006), lexical retrieval (Krauss, 1998), and the acquisition of reading and writing skills (Kiefer & Trumpp, 2012).

However, little evidence has revealed the role that body movements play in phonological learning (e.g., the perception and production of phonemes), especially in the context of second language learning. Therefore, the present dissertation aims at evaluating the potential benefits of using a variety of hand movements in the context of embodied pronunciation training. It thus falls within the framework of the Embodied Cognition paradigm and its application to embodied learning.

With the emergence of the Embodied Cognition paradigm, its application to learning processes and education, in particular, has been largely discussed. With respect to first language acquisition, it has been shown that children's development of concepts and language processes are influenced by their embodied experiences (Wellsby & Pexman, 2014). In a recent review article, Shapiro and Stolz (2019) offered a comprehensive review on how the Embodied Cognition paradigm can be significant to education. The authors highlighted one area where Embodied Cognition research has interesting educational implications, particularly the role gestures play in learning (see also Goldin-Meadow, 2011, for a review). First, learners' gestures can inform the instructors as to whether they have understood the information being taught. Evidence shows that the more accurate the learners' gestures during learning, the more likely they were to show better learning outcomes (Goldin-Meadow, 2011; Goldin-Meadow et al., 2009). Second, encouraging learners to actively produce gestures can enhance learning by shifting the cognitive load from verbal to visuospatial memory storage. For instance, Broaders et al. (2007) found that active gesture production during learning could trigger the awareness of unexpressed and implicit ideas and aided learning. Therefore, instructors should bear in mind the effectiveness of gestures.

In this connection, gestural research has offered a large amount of evidence in relation to the importance of bodily action during learning. The use of manual gestures has long been analyzed in cognition and language learning. Studies by Goldin-Meadow and colleagues hold that gestures not only reflect people's thoughts during verbal communication by providing information beyond verbal speech but also help to change people's knowledge by affecting both learners' cognition (effects on learners themselves) and their communication (effects on the learning environment) (e.g., Goldin-Meadow, 2010, 2011; Goldin-Meadow & Wagner, 2005). Accordingly, gestures could be regarded as a facilitator and a predictor for learning and thought (Goldin-Meadow, 2010, 2011). These findings confirm that gestures play an unneglectable role in learning, especially in children's first language acquisition (Goldin-Meadow, 2018; Iverson & Goldin-Meadow, 2005).

The present dissertation stems from the theoretical view supported by the Embodied Cognition paradigm and aims at evaluating the potential benefits of embodied pronunciation training, which involve a variety of visuospatial hand gestures and their effects on phonological learning. In the following section, we review the cognitive theories of gesture-speech integration that back up many of the claims made by the Embodied Cognition paradigm.

1.1.2 Cognitive theories of gesture-speech integration

During face-to-face communication, people gesture and gestures do not only affect the communication itself but also modulate the speakers' thoughts (Goldin-Meadow, 2010, 2011; Goldin-Meadow & Wagner, 2005). McNeill (1992) defined gestures as movements accompanying

speech, which are typically made by arms and hands and are simultaneously produced together with the speech flow (McNeill, 1992, p.11). However, gestures are not just a series of hand movements in space, but "symbols that exhibit meanings in their own right" (McNeill, 1992, p. 105). Thus, hand movements like self-touching and object manipulation are not considered gestures (McNeill, 1992, p. 78). Co-speech gestures can be categorized as: imagistic (e.g., iconic, metaphorical, deictic, that is, gestures which have a clear referential component) and non-imagistic (e.g., beat gestures, or gestures which do not represent referential meaning) (McNeill, 1992, p. 78)[1].

Several theories have been proposed to explain the role that gestures play in language and cognition. For instance, Krauss et al. (2000) proposed the "lexical gesture process model," suggesting that gestures can aid speech production by facilitating speakers at the stage of lexical retrieval. Recently, de Ruiter (2017) formulated the "Asymmetric Redundancy Hypothesis," claiming that iconic gestures provide additional visual and redundant information. This redundant information provides extra channels for speakers to correctly perceive and comprehend speech. Thus,

[1] McNeil (1992) defines the four types of gestures as follows: (1) A gesture is iconic if it bears a close formal relationship to the semantic content of speech (p. 78); (2) Metaphorical gestures are similar to iconics in that they present imagery, but present an image of an abstract concept (p. 80); (3) Deictic gestures are pointing movements, which are prototypically performed with pointing finger. (p.80); and (4) Beat gestures are defined as movements that do not present a discernible meaning, and they can be recognized positively in terms of their prototypical movement characteristics. (p.80)

gestures do not merely complement but enhance communication (de Ruiter et al., 2012).

Similarly, Kita and Özyürek (2003) claimed that since gestures come from the interface representation between speech and spatio-motoric processing (Interface Hypothesis), they contribute to speech planning specifically by helping speakers organize the spatial information in speech. Following this, Kita et al. (2017) extended the framework to the "Gesture-for-Conceptualization Hypothesis," which claims that all representational gestures contribute to speaking and thinking and schematize information that facilitates people's conceptualization. In this way, gestures facilitate speakers' speech production.

From a broader perspective, Hostetter and Alibali (2008) proposed the GSA framework (Gesture-as-Simulated-Action Framework) to account for how gesture is produced from an embodied cognitive system. According to GSA, human's perceived stimuli and the action taken due to the stimulation mutually determine each other, and the generation of gestures and actions share the same processing system. Therefore, gestures are viewed as elements that make cognition visible.

1.1.3 Gestures and cognitive load reduction

In the context of educational research, the Cognitive Load Theory (Chandler & Sweller, 1991; J. Sweller, 1988) proposed three types of cognitive demands during the learning process: (a) the intrinsic cognitive load, which is determined by the learning materials, (b) the extraneous cognitive load, which is due to the instructional design, and (c) the germane cognitive load, which reflects the efforts that learners made to the

learning (J. Sweller et al., 1998). Proper instructional design should reduce the extraneous cognitive load while increasing the germane load for the learners, especially when the learning outcome is largely limited by working memory, such as learning a foreign language (Paas & Sweller, 2012). Moreover, as Risko and Gilbert (2016) contend, people usually send the cognitive demands either "onto the body" (e.g., in order to see a rotated picture, one may tilt the head to normalize the orientation) or "into the world" (e.g., instead of remembering a phone number by the head, one can choose to write it on a paper and retrieve this information when needed) (p. 677). This is tightly linked to the Embodied Cognition paradigm, given that the interaction between mind, body, and environment shifts information to body movement so as to offload cognitive demands (Risko & Gilbert, 2016).

A number of studies have found that gestures tend to reduce cognitive load during learning. First, producing gestures could help to save cognitive resources to enhance memorization. For example, Goldin-Meadow et al. (2001) found that in a dual task (remembering letters or words while explaining math problems), people who were allowed to gesture could remember more items than those who did not. Second, gestures benefit speech production by reducing cognitive resources even when the referent is not visibly present. Ping & Goldin-Meadow (2010) showed that being allowed to make gestures can help children recall more words than not doing so, regardless of whether or not the object is present. Third, it is producing meaningful gestures rather than random hand movements that reduces demands on working memory. For instance, Cook et al. (2012) also proposed a dual task for three groups of participants to remember letter series while solving math problems. During the task, the

first group of participants had to produce gestures; the second group had to make meaningless hand movements; the third group was not allowed to move their hands. The results showed that participants could recall more items when producing meaningful hand gestures than producing random hand movements or keeping their hands still.

By contrast, some studies reported contradictory findings, suggesting that gestures may tax learners' cognition in some cases. Especially in mathematical instruction, gestures were found not to be always helpful during learning (Yeo et al., 2017) nor help retain the learning effects (Byrd et al., 2014). Yeo et al. (2017) asked a math teacher to teach linear equations to children aided with graphs. The teacher orally explained the equations while pointing to (a) the graphs, (b) the equations, (c) both graphs and equations, or (d) neither. However, pointing to the equations led to fewer learning outcomes than pointing to the graphs or with no gesture, although all the students showed substantial learning outcomes after the training. Similarly, Byrd et al. (2014) found that learning equations by performing hand gestures yielded similar improvement from pretest to immediate posttest, compared to non-gestural training methods. However, after four weeks, students who had received gestural training could not outperform those who had received non-gestural training, suggesting that gestures may not always make learning last.

One of the reasons that may account for the negative results is that when gestures provide redundant information, it may in turn moderate or even interfere with the effects of learning and retention (Byrd et al., 2014, p. 1986; Yeo et al., 2017, p. 9). Therefore, given the mixed findings in the previous studies, more work is needed to comprehensively assess the role of gestures in learning.

1.2 Gestures in second language learning

Research has shown that the use of gestures has been an essential component of second language acquisition. Gullberg (2006) provided a comprehensive summary of the reasons for investigating the connection between gesture and second language learning. She claimed that people acquire both language and gesture simultaneously, given that gestures are cross-cultural phenomena. Therefore, assessing learners' gestures can provide an insight into their learning process. More importantly, the use of gestures affects not only learners but also their interlocutors, highlighting the fact that hand gestures may aid comprehension and overall acquisition. In what follows, we will summarize empirical work on the beneficial role of gestures in L2 learning, with a focus on vocabulary and pronunciation learning.

1.2.1 Effects of gestures for vocabulary learning

Experimental and classroom research has evaluated the effects of representational gestures (i.e., gestures depicting their referents, e.g., metaphoric and iconic gestures) on L2 vocabulary learning from a variety of aspects. Allen (1995) trained 112 English-speaking adults to learn 10 French expressions either with or without gestures. Her results showed that students who learned the target expressions accompanied by gestures had a greater immediate recall and a smaller decay in recall after two months than those who learned them without gestures. Tellier (2008) found similar results with children by showing that children performing iconic gestures better recalled the target words than those who learned new words by viewing pictures. Later, in order to investigate whether the congruency between gestures and the semantic meaning affects word

learning, Kelly et al. (2009) trained adult learners of Japanese to learn Japanese verbs by observing (a) speech only, (b) speech and congruent iconic gesture, (c) speech and incongruent iconic gesture and (d) repeated speech. The results showed that participants recalled the largest number of words that were trained with congruent gestures and the least number of words that were accompanied by incongruent gestures, suggesting that gestures should congruently encode the semantic meaning of the word to be learned. Some later studies expanded the conclusion by showing that even though iconic gestures are not typically related to the word meaning, as long as they can be idiosyncratically mapped to the meaning, learners still benefit from observing them (Huang et al., 2019) and that observing hand gestures favors L2 vocabulary learning only when the phonetic demands are not very high (Kelly & Lee, 2012).

In line with these findings, Macedonia and colleagues reported that actively producing representational gestures not only helped participants to learn foreign words with concrete meaning (Macedonia et al., 2011) and abstract meaning (Macedonia & Knösche, 2011) but also led to better accessibility of newly learned words in memory when creating new sentences (Macedonia & Knösche, 2011). Then, Krönke et al. (2013) confirmed that actively performing meaningful hand gestures as opposed to random body movements yielded deeper semantic encoding of novel words. Morett (2014) further demonstrated that hand gestures may facilitate three cognitive processes during L2 word learning, namely, communication, encoding, and recall. Later, Macedonia and Klimesch (2014) conducted a fourteen-month classroom study and found that producing iconic gestures significantly enhanced vocabulary learning in the long

term. Moreover, gestures should be produced spontaneously in conjunction with speech so as to aid L2 word learning, but this facilitative role of gesture tends to be type-specific (i.e., deictic gestures) (Morett, 2018). Finally, in a recent study, N. Sweller et al. (2020) found that in learning L2 words and maintaining the memorization of the meaning, both observing and producing iconic hand gestures was equally effective.

Some studies have assessed the role of gestures that do not represent referential meaning (e.g., beat gestures) in the memorization and the acquisition of L2 vocabulary. So et al. (2012) compared the different roles that observing representational and beat gestures play in memorization with both children and adults. They asked participants to remember words with (a) iconic hand gestures depicting the semantic meaning of the words; (b) beat gestures; (c) or no gesture. The results revealed that when recalling the words, adults benefited from iconic and beat gestures equally, but only iconic hand gestures helped children to better recall the verbs. The results pointed to the fact that if non-representational gestures are to be used to enhance memorization, they will work better with adults. Applying this line of research to L2 learning, Kushch et al. (2018) taught Russian words to Catalan-speaking adults under four conditions. In condition 1, neither prosodic nor visual prominence was made; in condition 2, both prosodic and visual prominence was presented; in condition 3, only the prosodic prominence was shown; and in condition 4, only the visual prominence was presented. The prosodic prominence was highlighted by an L+H* pitch pattern, while the visual prominence, by beat gestures. The results revealed that participants benefited the most from

the combination of beat gesture and prosodic prominence, which highlighted the fact that beat gestures could help the memorization of L2 words.

Similarly, hand gestures encoding phonological information (e.g., pitch information) have been shown to help learners in learning L2 words. For example, Baills et al. (2019) and Morett and Chang (2015) found that pitch gestures depicting lexical tones in space could help learners to memorize L2 words contrasting in lexical tones.

1.2.2 Effects of gestures for pronunciation learning

Turning to the learning of pronunciation in the L2 context, evidence is mounting that the use of hand gestures may also play a role in this field. In the present dissertation, we will use the term *visuospatial hand gestures* to refer to a variety of instructor's hand configurations that visually encode specific phonetic and prosodic properties of speech, including pitch, durational, articulatory, phrase-level prosodic features, etc. The gestures in the present dissertation are instructional and mainly encode phonetic and prosodic features, and we thus term them as *visuospatial hand gestures*. According to the specific features that visuospatial hand gestures depict, they can further be classified as (a) pitch gestures (e.g., gestures mimicking *F0* movements), (b) durational gestures (e.g., gestures showing phonemic contrasts in duration), (c) gestures encoding articulatory features (e.g., gestures cueing certain segmental features, such as aspiration contrasts of consonants, etc.) and (d) prosodic gestures (e.g., gestures mimicking prosodic features, like pitch and duration, at the phrase-level). It is important to note that these hand gestures are not

strictly part of McNeill's (1992) classification of communicative gestures.

The upcoming review of the literature will show some inconsistent results on the role of using hand gestures for pronunciation training. For example, while pitch gestures have been demonstrated to facilitate the perception of L2 pitch features, durational gestures have not been found to have such benefits on perception, but they do facilitate production. Moreover, only a couple of studies have been conducted on gestures that encode articulatory information, and even fewer studies focused on the role of gestures that encode phrase-level prosodic features. As we will see, the mixed results from the empirical studies suggest that further studies should be conducted to fill this gap.

a) Effects of beat gestures

Beat gestures are associated with prosodically prominent positions and thus can serve as highlighters of rhythm. Krahmer and Swerts (2007) found that visual beats had a similar function to the pitch accent when making emphasis, and if the speakers produced a visual beat on the prominent word, the prominence would be acoustically perceived as stronger. Moreover, compared to other types of visual beats (e.g., rapid eyebrow movement), hand gestures can lead the speech addressees to perceive the corresponding words as more prominent.

In the field of L2 pronunciation, Gluhareva and Prieto (2017) investigated the effects of beat gestures on pronunciation on the discourse level. Twenty Catalan learners of English were trained with videos, where half of the videos presented beat gestures to speech prominence while the

other half did not. They then responded to some given contextual prompts. The results showed beat gestures could reduce learners' foreign accent in more difficult discourse items. Kushch (2018) also recruited Catalan learners of English to perform gestures on the same set of contexts used by Gluhareva and Prieto (2017). But they were trained by either performing or observing beat gestures. The results showed imitating beat gestures yielded more gains than observing them. Another study by Llanes-Coromina et al. (2018) further demonstrated that actively producing beat gestures could help Catalan learners of English to achieve better accentedness, comprehensibility, and fluency scores compared to merely observing them.

By contrast, a recent study found that beat gestures or gestures encoding durational features did not facilitate the production of L2 lexical stress. In a between-subject study, van Maastricht et al. (2019) trained Dutch speakers to learn Spanish lexical stress produced with (a) hand gestures mimicking the enhanced duration of stressed syllables by moving both hands to the side of her body, (b) beat gestures stroke to the stressed syllables, and (c) no gestures. However, all three groups of participants showed similar gains in their production accuracy of Spanish lexical stress, suggesting that neither gesture played a facilitative role in learning lexical stress.

In sum, although with mixed results, it seems that beat gestures could boost the L2 pronunciation learning at least regarding rhythmic features.

b) Effects of pitch gestures

Regarding the value of pitch gestures (or hand movements mimicking melodic patterns in speech), a handful of studies have shown that they can boost learning of novel tonal and intonational features in a second language.

First, a number of studies have demonstrated that pitch gestures significantly improved the perception of L2 lexical tones. For instance, Morett and Chang (2015) taught English speakers to learn Mandarin words by showing them (a) pitch gestures depicting the tonal patterns, (b) iconic gestures showing the meaning, or (c) no gestures. They found that pitch gestures could strengthen the relationship between lexical meaning and tones. Hannah et al. (2017) investigated the relationship between pitch gestures and the perception of Mandarin Chinese tones. They asked English speakers to identify the Mandarin tones and found that facial and gestural information lent a hand to the perception of novel tones and that when perceiving tonal features, learners of Chinese used a multimodal strategy that relies on both acoustic and visual tonal cues. Moreover, Baills et al. (2019) confirmed that both observing and producing pitch gestures was favorable to the perception of L2 tonal patterns as well as the learning of words contrasting in tones. Furthermore, Zhen et al. (2019) examined the role of pitch gestures on perceiving lexical tones varying in a set of parameters: congruency (whether gestures moving in the congruent direction to that of the pitch), modality (viewing or performing gestures), and spatial domain of gesture movement (whether gestures performed horizontally or vertically). They found that gesture observation and production had equal benefits on the perception of lexical tones, as long as they congruently encoded the pitch track. However, when the

gestures were shown in a horizontal panel, performing hand gestures was proven better than observing them. This conclusion highlights the importance of cross-modal learning of L2 pitch features as well as the importance of the adequacy in the visual presentation of the target gestures.

Second, some studies have explored the effects of pitch gestures on the production of L2 lexical tones. For instance, in a classroom setting, Chen (2013) taught L2 Chinese learners from different countries to learn Chinese tones with or without pitch gestures. They found that the experimental group outperformed the control group on tonal production accuracy scores and the accuracy of responses to the teachers' queries on tones. By contrast, another study found that observing pitch gestures encoding Mandarin tonal patterns only played a moderate role in the simultaneous speech imitation: it only helped the tonal accuracy of the falling tone when participants were asked to imitate the lexical tones after a native Chinese speaker (Zheng et al., 2018).

Apart from lexical tones, Ghaemi and Rafi (2018) compared the effects of printed visual stimuli and pitch gestures on the learning of English stress patterns. While participants in one group learned the stress patterns by seeing words printed on a piece of paper and repeatedly hearing them spoken aloud, participants in a second group learned them in the same way except that stressed syllables were written in boldface. By contrast, the third group was shown stressed syllables in boldface and also co-speech hand gestures made to the stressed syllables. The gesture was a forward, horizontal hand movement during unstressed syllables and upward hand movements, which mimicked the speaker's pitch rise when producing stressed syllables. The results revealed that although all three groups showed an improvement at delayed posttest (two weeks after the

training), the training with pitch gestures yielded the best learning outcome.

The benefits of pitch gestures have also been documented at the sentence level. Kelly et al. (2017) reported that gestures signaling the pitch features of Japanese yes/no questions (an upward hand movement) and affirmative questions (a downward hand movement) helped learners make intonational distinctions. Yuan et al. (2019) confirmed the beneficial effects of pitch gestures on the learning of L2 intonation. They trained Mandarin-speakers with basic Spanish proficiency to learn Spanish intonation patterns, namely statements, yes-no questions, and requests. Half of the participants were trained by observing speech and gestures performed over the nuclear configuration (e.g., an upward gesture for rising tong), while the other half, by observing speech only. Participants exposed to gestural training improved their realization of the intonation patterns better than the other group, suggesting that observing hand gestures depicting nuclear intonation contours can favor the learning of L2 intonational patterns at the phrasal level.

Finally, some studies have shown that the perception of acoustic pitch and hand movement in space share common representational and processing resources. Casasanto et al. (2003) showed lines prolonging vertically (bottom to top) and horizontally (left to right) to two groups of participants respectively and asked them to reproduce either stimulus displacement or stimulus pitch. The results showed that vertical displacement strongly modulated participants' estimates of acoustic pitch inputs, but horizontal displacement did not, suggesting that there is a linguistically and conceptually metaphoric relationship between space and pitch.

Connell et al. (2013) investigated the role visual movements play in perceiving pitch and found that when watching upward or downward gestures, people tended to perceive the pitches higher or lower than they actually were, supporting the 'shared representation' explanation for the relationship between pitch and space. This was further supported by means of neurophysiologic measures, where it was found that judging the auditory stimuli activated unimodal visual areas of the brain, which means there is an overlap between the processing of auditory pitch height and visuospatial height in the visual brain area (Dolscheid et al., 2014).

c) Effects of durational hand gestures

Recent studies using durational gestures (e.g., hand movements cueing phonological length contrasts) to boost the perceptual processing of Japanese durational vowel contrasts have yielded mixed results. First, Hirata and Kelly (2010) reported that observing a beat gesture representing the short vowel in combination with a hand sweep for the long vowel did not show positive effects on the perception of durational contrasts. Later, Hirata et al. (2014) compared the effects of observing and producing syllable gestures (a hand sweeping for a long vowel and a beat for a short vowel) and mora gestures (two beats for a long vowel and one beat for a short vowel). However, only observing syllable gestures was effective in the perception of the durational contrasts in the most balanced way between the word-initial and word-final position as well as at both fast and slow speech rates. By contrast, mora gestures did not facilitate learning. In addition to the auditory learning, the participants were also taught vocabulary items bearing the vowel-length contrasts using the same syllable and mora gestures used in Hirata et al. (2014). However, participants had similar outcomes in vocabulary learning with either observing or

producing those gestures. In a subsequent experiment, Kelly et al. (2017) found that observing hand gestures representing durational contrasts still did not help learners to hear differences in vowel length. These findings were further supported by electrophysiological evidence (Kelly & Hirata, 2017). Taken together, the experiments carried out by Hirata, Kelly, and colleagues suggested that neither observing nor actively producing hand gestures signaling vowel-length distinctions facilitates the perception of durational contrasts.

The authors thus concluded that hand gestures had only limited effects on the perception of durational contrasts. However, using hand gestures to facilitate the learning of durational features is constantly reported by classroom observations (Hudson, 2011) and suggested by teaching proposals (Roberge et al., 1996). In these reports and suggestions, the teachers do not make use of beat but try to illustrate the durational contrast by horizontal hand movement (Roberge et al., 1996) or by moving both hands horizontally outward to show a long vowel while by approaching two fingers together to show the short vowel (Hudson, 2011). In our view, the negative results obtained in some of the abovementioned studies may have been due to methodological reasons. First, as the authors themselves suggest (Hirata & Kelly, 2010: 306; Hirata et al., 2014: 9), "there is evidence that layering too much multimodal information onto novel speech sounds may overload the system and actually produce decrements in perception and learning." For example, Hirata and Kelly (2010) showed that while English learners benefited from seeing lip movements to distinguish Japanese long and short vowels, adding hand gestures to lip and audio training actually canceled the positive effects of the lip in-

formation. Second, the use of the contrasting pair of beat gesture/sweeping gesture as mimicking short/long vowel distinctions might not be effective for listeners. 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 themselves 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). 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 in mimicking a vowel-length difference.

Interestingly, there is behavioral evidence linking horizontal movements with the mental representation of duration. Casasanto and Boroditsky (2008) reported a series of experiments that showed that spatial movement strongly modulated people's estimation of temporal duration. Later, Cai and Connell (2012) found that time and space are tightly linked to each other, and the relationship between temporal duration and spatial duration was relative to the modality of perception. That is, when people perceive spatial length through both tactile and visual modalities, the perception of spatial duration strongly affects their estimation of temporal duration. Furthermore, Cai et al. (2013) found that hand gestures moving in space can significantly modulate people's estimation of temporal duration. More specifically, participants were asked to listen to musical notes accompanied by horizontal hand sweep gestures moving in long or short distance and then reproduce the temporal duration of each note by pressing a button based on their subjective estimation. It turned out that a note would be estimated longer if it was accompanied by long moving gestures and shorter if the accompanying gesture moved shorter in space.

These behavioral studies suggest that the durational contrasts in speech should be represented by contrasting horizontal hand movements.

Based on the evidence mentioned above, we believe that more empirical evidence is needed to demonstrate the potential benefits of the durational hand gestures on the phonological learning of vowel-length contrasts. To fill this research gap, Study 1 aims to further investigate the role of hand gestures encoding durational information by reshaping the manual configuration into a horizontal hand sweep gesture and by taking into account its effects not only on perception but also on production.

d) Effects of gestures encoding articulatory features

Even though L2 pronunciation teaching practices suggest that a variety of useful hand shapes and hand movements are used by instructors in their L2 classrooms (Hudson, 2011; Smotrova, 2017; Y. Zhang, 2002), little experimental work has been conducted to assess the role of hand gestures mimicking specific articulatory features of segments (e.g., gestures encoding spatiotemporal parameters such as holding fingers and thumb together and separating them quickly to cue /p/).

To our knowledge, only three experimental studies have tested the potential benefits of *observing* hand gestures cueing segmental features on L2 pronunciation learning, two of them dealing with aspiration features (Amand & Touhami, 2016; Xi et al., 2020) and the other one dealing with labiodental consonantal features and rounded vocalic features (Hoetjes & van Maastricht, 2020).

Amand and Touhami (2016) explored the effects of gestures in facilitating French learners of English to pronounce the English word-final plosives. The results showed that, although participants generally improved in the production of unreleased stops, training with hand gestures yielded significantly more improvement than without hand gestures. This study demonstrated that gestures boosted the learner's awareness about aspiration patterns, which helped them in properly producing the word-final plosives.

Hoetjes and van Maastricht (2020) compared the effects of pointing gestures and iconic gestures on the learning of L2 Spanish /u/ and /θ/. The results revealed that /u/ was easier than /θ/ in acquisition. For /θ/, the pointing gesture appeared to be helpful, while for /u/, training with an iconic gesture (i.e., rounding the palm to indicate the rounding of the lips) was proven beneficial. Interestingly, the iconic gestures were particularly helpful for the learning of /u/, while harmful for /θ/. This study thus presents an interesting interaction between the complexity of gesture shape and phoneme-to-be-learned. That is, gestures used to train L2 phonemes should be adequate, and for an L2 phoneme with greater difficulty, the gestures for instruction should not be complex so as to offload the cognitive demands for participants.

More recently, Xi et al. (2020) trained 50 Catalan speakers to learn six pairs of Mandarin consonants with or without hand gestures. Three pairs were plosives /p-pʰ, t-tʰ, k-kʰ/ which contrast in aspiration and differ in the absence/presence of a strong air burst; the other three pairs were affricates /ts-tsʰ, tɕ-tɕʰ, tʂ-tʂʰ/ which are phonologically described as unaspirated-aspirated contrast but also acoustically differ in the duration of frication period. They used a fist-to-open-hand gesture to cue the air

burst of the aspirated consonants. The results revealed that while observing this gesture significantly improved participants' pronunciation of plosive pairs, it failed to aid the pronunciation of affricates. This is because the gesture, which mimicked a strong airburst for aspiration might be inadequate to cue the affricates /tsʰ, tɕʰ, tʂʰ/ since they have a longer frication period than their counterparts /ts, tɕ, tʂ/. This study not only added new evidence to support the positive role of hand gestures in L2 pronunciation learning but also emphasized that gestures should adequately mimic the target features.

To summarize, although previous studies have shown that *observing* hand configurations and movements encoding a variety of phonetic features (lip rounding, tongue position, aspiration, etc.) showed beneficial effects on the learning of L2 segmental features, it is not yet clear the role of *producing* hand gestures. Moreover, as shown by Hoetjes and van Maastricht (2020) and Xi et al. (2020), participants should *observe* adequate hand gestures during the training to achieve an improvement in producing the target phonemes. It is, therefore, crucial to assess the appropriateness of learners' gesture performance when they *produce* the target gesture during training. Finally, all the above-mentioned three studies (Amand & Touhami, 2016; Hoetjes & van Maastricht, 2020; Xi et al., 2020) did not take into account the potential effects of these gestures in differentiating word meaning. More importantly, most of the previous research (Baills et al., 2019; Kushch et al., 2018; Morett & Chang, 2015) in relation to the effects of non-representational gestures on the memorization of L2 words mainly tested its immediate effects.

Therefore, a second study was proposed (Study 2) to test the potential benefits of producing visuospatial hand gestures in the learning of L2

segmental features. Following Xi et al. (2020), we selected Mandarin Chinese as the target L2 and L1 Catalan speakers as our subjects, given that this Romance language does not have aspiration contrasts in plosives (Wheeler, 2005). The target phonemes were the plosive consonants contrasting in aspiration /p-p^{h}, t-t^{h}, k-k^{h}/. The novelty of Study 2 is the fact that we assessed the role of gesture production in learning segments as compared to previous studies, which only involved gesture observation (Amand & Touhami, 2016; Hoetjes & van Maastricht, 2020; Xi et al., 2020). Importantly, the study takes into account the appropriateness of learners' gesture performance during the training, which is coupled with the congruency between gesture shape and the phonetic feature it attempts to represent (Xi et al., 2020). Finally, the delayed effects of the embodied training involving gestures cueing segmental features are assessed.

1.3 Embodied prosodic training: The effects of integrating gestures and prosody for second language pronunciation learning

When assessing oral proficiency in a second language, comprehensibility, accentedness, and fluency are the most commonly used measures (Munro & Derwing, 2015). Although comprehensibility is essentially affected by grammar, lexis, and discourse complexity (Isaacs & Trofimovich, 2012; K. Saito et al., 2016, 2017; Trofimovich & Isaacs, 2012), pronunciation components also play an essential role, from a wide range of suprasegmental features (Crowther et al., 2016; Isaacs & Trofimovich, 2012; Munro & Derwing, 2001; K. Saito et al., 2016, 2017; Trofimovich & Isaacs, 2012) to segments with high functional load

(Munro & Derwing, 2006; Suzukida & Saito, 2019). In addition, fluency measures were also found to contribute to comprehensibility (Crowther et al., 2016; Isaacs & Trofimovich, 2012; K. Saito et al., 2017).

Unlike comprehensibility, accentedness is primarily related to pronunciation measures (K. Saito et al., 2016). Many studies suggest that suprasegmental features seem to weigh more in the perception of foreign accentedness (Anderson-Hsieh et al., 1992; Boula de Mareüil & Vieru-Dimulescu, 2006; Trofimovich & Baker, 2006); while others consider segmental accuracy a vital cue for native judge of accentedness (Rognoni & Busà, 2014; K. Saito et al., 2016, 2017; Trofimovich & Isaacs, 2012).

Despite the long-standing debate on which factors should be prioritized in teaching practice, recent meta-analysis and reviews suggest that both suprasegmental and segmental features should be trained during pronunciation instruction (J. Lee et al., 2015) and that teachers should take advantages of the interactions between the two (X. Wang, 2020). Therefore it is effective to organize pronunciation training under the prosodic structure of the target language, which combines both the suprasegmental and segmental components and their interactions (Zielinski, 2015).

1.3.1 Effects of prosodic training on global pronunciation, suprasegmental and segmental features in an L2

Research in second language pronunciation has made attempts to implement prosodic training (i.e., implicit training focusing on the prosodic form of an L2) in teaching practice, which aims at improving learners' global pronunciation proficiency. Derwing et al. (1998) found that focusing on segments or prosody revealed positive effects on controlled

and spontaneous speech production, but only prosodic training improved comprehensibility and fluency in spontaneous speech, pointing to the needs for global pronunciation training in L2 teaching practice. Along this line, recent studies consistently found that L2 learners benefited from prosodic training in improving global pronunciation proficiency. For instance, Gordon et al. (2013) and Gordon and Darcy (2016) found that explicit training on suprasegmental features enhanced learners' comprehensibility while training on individual segments like vowels did not. Similarly, Saito and Saito (2017) confirmed that training on supra-segmental features improved learners' comprehensibility and helped learners acquire correct intonational patterns. In a recent study, R. Zhang and Yuan (2020) showed that while both segmental and prosodic training yielded significant gains in participants' pronunciation, only prosodic training improved comprehensibility in spontaneous speech production and helped maintain these gains at the delayed posttest.

In training prosody, another line of research is computer-assisted pronunciation training, where much attention has been paid to the training of intonation. The key idea of this training paradigm is to provide learners with visual representations of the target pitch contours created by computer software and their own pitch contours during the prosodic training. Learners therefore can visually capture the differences in intonation between their speech and the model speech (Olson, 2014). A series of early studies reported that learners provided with visualized pitch contours on the computer significantly improved their global pronunciation (de Bot, 1983; Weltens & de Bot, 1984a, 1984b). Hardison (2004) employed a similar training method and found that training intonation aided by computer software was beneficial in reducing learners' foreign

accents. In line with these studies, it was confirmed that visually highlighting the intonation, stress and the waveform of the target language improved learners' global oral proficiency (Gorjian et al., 2013), intonation (Hincks & Edlund, 2009; Ramírez Verdugo, 2006) as well as the accuracy of stress patterns in both controlled (Tanner & Landon, 2009) and spontaneous speech production (Schwab & Goldman, 2018). In addition, Wang et al. (2016) found that using computer software casting linguistic rhythm into musical rhythm helped the acquisition of rhythmic patterns of the target language.

By contrast, only a handful of studies explored the potential effects of prosodic training on the learning of segmental features. Missaglia (2007) argued that the vocalic and consonantal mispronunciations in an L2 are not likely due to inaccuracy in the production of a single segment but rather due to insufficient suprasegmental competence (Missaglia, 2007, p. 239). She proposed to train Italian speakers on the suprasegmental features of German, targeting the accenting and de-accenting patterns. During training, learners were asked to exaggeratedly produce only one stressed syllable in each sentence, and, accordingly, the rest of the syllables in the sentence were reduced. Therefore, vowel reduction and centralization naturally occurred. Missaglia (2007) trained Italian learners with beginning-level German proficiency in two groups: One focusing on segmental training and the other group, prosodic training. The results showed that prosodic training triggered more improvement in global pronunciation and fewer segmental errors than segmental training. This finding validated the author's assumption.

Likewise, Saito and Saito's (2017) study showed that L2 learners of English who had received prosodic training showed improvements in the

pronunciation of segments. They produced longer and clearer vowels in stressed syllables and showed more proper vowel reduction in stressed unstressed syllables. The authors claimed that the improvements in segmental production stemmed from the gains in rhythmic structures. Moreover, Hardison (2004) also found that the benefits of prosodic training generalized to the improvement of segmental accuracy. She trained 16 English-speaking learners in French prosody with computerized visual feedback of pitch contours. The results revealed that participants' prosodic accuracy was improved after training and this improvement was generalized to segmental accuracy as well.

Nevertheless, some studies failed to find positive effects of prosodic training on the learning of L2 segmental accuracy. In Gordon and Darcy (2016), for example, although segmental training on specific vowels tended to improve their pronunciation of the trained vowels with small effect size, it failed to improve learners' comprehensibility. By contrast, while prosodic training did not seem to work on segments, it did facilitate learners' comprehensibility. This finding appeals to the need for prosodic training in L2 teaching practice rather than merely focusing on specific phonemes. However, more evidence is needed for assessing the role of prosodic training in learning segments.

In sum, prosodic training, in general, plays a positive role in improving learners' global pronunciation and suprasegmental features. However, only a handful of studies have investigated the role of prosodic training in the learning of L2 segments, with mixed results (e.g., Gordon & Darcy, 2016; Missaglia, 2007, 1999; Saito & Saito, 2017). Importantly, researchers have also noted the need to assess prosodic training at the discourse level (Levis & Pickering, 2004; Seferoğlu, 2005). Many of the

previous findings measured the learning outcome at the word level or the sentence level. Therefore, it is necessary to evaluate the effects of prosodic training at the discourse level, which is one of the novelties of Study 3.

1.3.2 The interaction between prosodic structure and the pronunciation of segments

Then, why would prosodic training play a role in the improvement of segmental accuracy? First, we claim that prosodic and segmental features both contribute to speech perception and production, and they are interdependent to each other and function in an integrated manner. Second, prosody can play a bootstrapping role in language acquisition. Following the Prosodic Bootstrapping hypothesis, paying attention to prosodic features of the target language like pitch and rhythm may help the learning of lexis and syntax for children in their native language (Christophe et al., 1997, 2008). This bootstrapping effect not only plays a role in early first language acquisition of typically developed children but also shows potential applications for speech therapy (Bedore & Leonard, 1995). Some recent empirical studies have extended this hypothesis to the field of L2 acquisition, where illustrating speech rhythm of an L2 to children may improve their sentence imitation abilities (Campfield & Murphy, 2014). Given the intricate play between prosodic and segmental structure, we believe that the Prosodic Bootstrapping hypothesis can be extended to the learning of L2 phonological learning, whereby the prosodic structure can be used to bootstrap the pronunciation of specific segmental features.

Moreover, two complementary proposals capture the interaction between prosodic prominence and enhanced segmental articulation. On the one hand, the Sonority Expansion hypothesis (e.g., Beckman et al., 1992) predicts that vowels in the speech prominent position may be produced with the jaw more opened along with lingual backness. Therefore, more energy is released from the mouth, resulting in an enhanced first formant value (F1). On the other hand, the Hyperarticulation hypothesis (de Jong, 1995) holds that speech prominence may trigger enhanced articulation of lip roundedness and backness in vowels. While the two hypotheses are somewhat compelling in that the latter one predicts that stressed vowels are not only distinct from unstressed vowels in sonority (shown by F1) but also in non-sonority (e.g., vowel backness measured by F2), they both provide theoretical assumptions for the interaction between prosody and segments, especially in the production of vowels. By contrast, vowels produced in a non-prominent or unstressed position may undergo compression in articulation and acoustic features (Walker, 2011, p. 16), resulting in a shorter duration, lower F1 (especially in non-high vowels), and reduced vowel space (Herrick, 2003; Lindblom, 1963; Padgett & Tabain, 2005).

Empirical evidence has revealed that the realization of segments like vowels is largely affected by their prosodic position. For example, in spontaneous speech, English vowels in prominent positions may be enhanced in sonority (higher values of the first formant), and the front vowels tend to be hyperarticulated (indicated by higher values of the second formant) (Mo et al., 2009). Cross-linguistic evidence also noted that the distinctiveness of vowels is enhanced in focus positions than in non-focus positions (Hay et al., 2006).

Moreover, the realization of segments may be influenced by emotional expressions (Estrada Medina, 2004, 2007). Particularly in French, utterances of *surprisal* differ from utterances of affirmation, as *surprisal* is marked by a prolonged and high-pitch sentence-final syllable and by accentuation in predicates. Therefore, the *surprisal* may affect the realization of duration, melodic structure, rhythmic structure, and segmentation in both L1 and L2 speech production (Estrada Medina, 2004). Based on this observation, a pilot study analyzed the L2 speech production of French by four Spanish-speakers and found that when the front vowels occurred in the *surprise* utterances, learners would produce them more clearly and more nativelike (Estrada Medina, 2007).

Other studies on French have noted an interaction between prosody and vowel quality not only in L1 (Georgeton & Fougeron, 2014) but also in L2 speech (Santiago, 2021; Santiago & Mairano, 2019). Georgeton and Fougeron (2014) investigated whether the initial position of an Intonational Phrase has an effect on the articulation of vowels in native French speech. They found that when the vowels receive initial strengthening, they tend to be articulated with larger mouth aperture and lip width. As for tongue position, the initial strengthening effect is particularly strong: front vowels are realized with more front and back vowels tend to be more backward. Their findings underscored the importance of prosodic position on the realization of vowel quality. In terms of second language speech, a recent study suggested that the vowel space of French in L2 speech is expanded by strong prosodic positions (Santiago & Mairano, 2019). Specifically, vowels at the final position of an Intonation Phrase or at the edge of an Accentual Phrase had expanded vowel space and

longer duration than at word-internal non-accented positions, which suggests that strong prosodic positions may lead L2 learners to produce enhanced vowels. More recently, Santiago (2021) compared the effects of prosodic position on the realization of rounded-unrounded vowel contrasts (i.e., /i, e, ɛ/ vs. /y, ø, œ/) in native French speech and L2 French speech. He found that prosodic position has significant effects on the realization of rounded-unrounded contrast. Specifically, for both L1 and L2 speakers, the distinction between front rounded and unrounded vowels is enlarged in prosodically accented positions (both at initial or final positions of an Accentual Phrase) than in unaccented positions.

Briefly, it seems that the realization of vowel quality is largely affected by the prosodic structures in both L1 and L2 speech. This reinforces the claim that prosody and segments should be jointly introduced into the L2 classroom teaching (X. Wang, 2020) and that taking advantage of their interaction may achieve better training outcomes (Zielinski, 2015).

1.3.3 Verbotonal method as an embodied prosodic training approach

Given the findings on the positive role of embodied training (mainly via gestures) in L2 pronunciation, it is worth exploring whether embodied pronunciation training that includes phrase-level prosodic features would have more benefits than non-embodied approaches. In the context of the verbotonal method, the training of prosodic features like rhythm, accentuation, and intonation is prioritized. The method also recommends combining prosody and body movements like hand gestures for phonetic corrections at both the segmental and suprasegmental levels (e.g., Guberina, 2008; Intravaia, 2000; Renard, 1989).

A representative example of the techniques used by the verbotonal method is the phonetic correction of the French front rounded /y/, which is often mispronounced /u/ by Spanish speakers. The teacher is advised to place the vowel in rising intonation contexts so as to trigger a more target-like pronunciation (Renard, 2002). At the same time, adding an upward hand gesture while producing the rising intonation may be of help as well (Billières, 2002). In teaching practice, teachers can place the target vowel /y/ in various contexts with different intonational structures to create meaningful discourses (see Wlomainck, 2002, p.159). In addition, the pronunciation of a vowel is clearer and more intelligible when it bears an accent. Therefore, it is suggested to correct the mispronounced vowels in stressed syllables (Renard, 2002), which can also be highlighted by hand gestures.

Within this framework, a number of empirical studies have assessed the efficacy of the verbotonal method in actual teaching practice. Alazard et al. (2010) reported a pilot study that showed that the verbotonal method improved L2 French learners' fluency in a reading task. Later, Alazard (2013) extended the pilot study to an eight-week training experiment with beginner and advanced learners of French. She compared the verbotonal method to the articulatory method, a method that explicitly trains L2 segmental pronunciationL2 segmental pronunciation. The results showed that beginner learners of French improved their fluency in a reading task with the aid of the verbotonal method after four weeks of training, although the improvement was not maintained after eight weeks. A possible reason was that the introduction of reading activities may have had a negative impact on the training effects.

Alongside the research line on the verbotonal method *per se*, some studies have tried to integrate this method with other approaches. F. Z. Zhang (2006) proposed a multisensory approach based on the verbotonal system to teach Chinese prosody to Australian English speakers. She integrated the communicative approach and body movement to aid learning. Students who were encouraged to make use of body movements during the learning process were found to be more proactive and motivated than those who were trained in the traditional communicative approach. Regarding their performance in pronunciation, compared to the traditional communicative approach, students who received embodied training produced higher mean F0 value, wider pitch range, and more accurate tonal patterns. This study supported the role that body movements play in L2 prosodic learning. He et al. (2015) explored the possible integration of computer-assisted language learning and the verbotonal method. She trained Chinese undergraduate students to improve English pronunciation in two groups. The control group merely repeated English sentences after the teacher. Yet, the experimental group listened to sentences that only presented their rhythm and melody, with all the vowels and consonants removed by a low-pass filter. In this way, students could pay more attention to the prosodic features. Moreover, students in the experimental group were encouraged to perform body movements, like hand-clapping, making beat gestures to the rhythm, walking along with the melody, and stepping to the stressed syllables, all while listening to the filtered sentences. These classroom activities were assisted by an online computer system for students to record and compare their pronunciation to native speech. The results showed that students in the experimental group had more improvement in their speech than those in the control group, especially in terms of pronunciation, comprehensibility, and fluency. Later,

adopting He et al.'s (2015) method, Yang (2016) conducted a similar experiment with primary school children and obtained similar results. These results show that training prosody can be advantageous even for young learners. The prosodic and gestural highlighting strategies involved in the training phase offered the learners a reliable basis for correctly producing the target language.

However, in a recent pilot study with eight English-speaking learners of French, Alazard-Guiu et al. (2018) could not find positive effects of the verbotonal method on the improvement of segmental accuracy. They compared the effects of the verbotonal method and the articulatory method on the pronunciation of French vowels, but the results showed that only the F3 value of the /a/ sound was improved with both training methods. As this is a pilot study with a limited number of participants, further studies with a larger sample size are needed to back up the effects of embodied prosodic training on the pronunciation of L2 segments.

All in all, the abovementioned experimental studies successfully applied embodied prosodic training techniques (like the verbotonal method) to L2 pronunciation learning. However, an open question still remains, namely whether embodied prosodic interventions may benefit the pronunciation of L2 segments, as previous studies revealed mixed results. The main goal of Study 3 was to assess the effects of embodied prosodic training not only on global pronunciation proficiency but also on segmental accuracy in an L2. Importantly, the evidence in favor of the interaction between prosody and vowel quality in both L1 and L2 French speech (e.g., Estrada Medina, 2004, 2007; Georgeton & Fougeron, 2014; Santiago, 2021; Santiago & Mairano, 2019) backs up the training tech-

niques of the verbotonal method and encourages the use of various prosodic structures in pronunciation training. Together with the role of various types of embodied techniques in highlighting the prosodic structures (e.g., Gluhareva & Prieto, 2017; Kushch, 2018; Llanes-Coromina et al., 2018; Yuan et al., 2019), we believe that an embodied prosodic training that highlights melodic and rhythmic features at the sentence level will be an effective tool to improve L2 pronunciation at both segmental and suprasegmental levels.

1.4 General objectives, research questions, and hypotheses

The present dissertation aims to empirically assess the potential benefits of visuospatial hand gestures encoding novel pronunciation properties in the context of a multimodal approach to L2 learning. As such, this dissertation is couched within the Embodied Cognition paradigm and its application to language learning.

Even though the Embodied Cognition paradigm supports the benefits of body movements on language comprehension and lexical processing (see section 1.1 for a review), little work has been carried out on the role it plays in L2 pronunciation learning, especially at the segmental level. The present dissertation includes three multimodal training studies which actively make use of visuospatial hand gestures to boost the acquisition of L2 phonological features.

The general hypothesis is that the use of visuospatial hand gestures cueing phonetic properties will facilitate L2 pronunciation at both segmental and suprasegmental levels. We have a number of reasons to hypothesize

that the phonological module of language is tightly connected to the use of hand gestures. First, in face-to-face human interaction, speech is typically accompanied by hand gestures, which play an important role in the development of speech, from the babbling stage onwards. Second, the transition from primarily manual to primarily vocal language is a gradual process during the first years of life. It has been shown that mother-child communication often involves the use of gestures (deictic, iconic, etc.) and that maternal gestural input is helpful to enhance children's vocabulary size (Iverson et al., 1999). Third, behavioral and neuroscientific results have shown that hand gestures and speech share complex neuroscientific interaction (Gentilucci, 2003; Gentilucci & Corballis, 2006; Rusiewicz et al., 2014; Rusiewicz & Rivera, 2017). Fourth, there is evidence that language is embodied, since when either speech or sensorimotor actions take place, both the language and motor areas in the brain are found to be activated (Desai et al., 2010; Pulvermüller et al., 2005; J. Yang & Shu, 2016), suggesting that there is a close link between body movement and language.

Therefore, one may expect that not only suprasegmental but also segmental learning could benefit from the use of hand gestures. In this context, the present PhD dissertation has three specific goals. The first goal is to further test the benefits of visuospatial hand gestures in the learning of novel vowel-length contrasts. The second goal is to explore the potential benefits of visuospatial hand gestures cueing aspiration properties in the learning of novel aspirated plosives. Finally, the third goal is to test whether gestures encoding prosodic features (melodic and durational) at phrase-level may improve both the global pronunciation and the accuracy of vowels.

We carried out three between-subject training studies with a pretest and posttest design to assess the role that visuospatial hand gestures play in the learning of (a) Japanese vowel-length contrasts, (b) Mandarin aspiration contrasts in plosives, and (c) French front rounded vowels, with each study addressing one aspect. In addition, global pronunciation was also measured in Studies 2-3. The main research question for each of the three studies is the following:

(1) Does training with visuospatial hand gestures encoding durational differences in vowels help improve the perception and production of novel vowel-length contrasts? (Study 1)

(2) Does training with visuospatial hand gestures cueing aspiration features of plosives help improve the perception and production of novel aspiration features, as well as the acquisition of novel words bearing these contrasts? (Study 2)

(3) Does training with visuospatial hand gestures depicting sentence-level prosodic features boost the pronunciation of novel vocalic features, as well as the global pronunciation proficiency? (Study 3)

The upcoming chapters will be organized into three separate studies (Studies 1, 2, and 3):

- Study 1 (Chapter 2) trains Catalan speakers without prior knowledge of Japanese in the pronunciation of L2 durational features in a laboratory setting. It assesses whether producing visuospatial hand gestures mimicking durational properties

through horizontal hand movements helps Catalan speakers with no knowledge of Japanese to identify and imitate long and short vowels. We have two main predictions. We predict that producing durational hand gestures may (a) enhance Catalan speakers' accuracy in identifying Japanese vowel-length contrast and (b) increase the ratio of long vowels to short vowels in Japanese speech production. In a between-subjects experiment with a pretest and 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. The Gesture condition produced each word while simultaneously mimicking the visuospatial hand gestures, while the No Gesture condition repeated the words orally without any gestural stimulation. Pretest and posttest consisted of the identical vowel-length identification and imitation tasks with the test words embedded in short sentences. The identification task was evaluated by means of accuracy score, while the imitation task was analyzed using acoustic measures, namely the duration ratio of long vowels to short vowels.

- Study 2 (Chapter 3) trains Catalan speakers without prior knowledge of Chinese in the pronunciation of L2 aspiration features in a laboratory setting. It examines the potential benefits of visuospatial hand gestures cueing aspirated features in learning non-native aspirated plosives, with a focus on the accuracy of participants' gesture performance accuracy. We predict that producing hand gestures will help Catalan speakers

without any knowledge of Mandarin to (a) better perceive and produce Mandarin aspirated consonants and (b) maintain the memorization of the newly learned Mandarin words baring this contrast. We additionally predict that (c) gesture performance accuracy would impact the learning outcome. Sixty-seven Catalan participants memorized and learned to pronounce novel Mandarin words containing non-native aspirated plosives, with or without performing hand gestures. They were tested on perception, production, and word-meaning recognition in a pretest, a posttest immediately after the training, and a delayed posttest after three days. The perception and the word-meaning recognition tasks were assessed by accuracy score, while the imitation task was assessed by measuring the voice onset time (VOT) of the aspirated plosives, plus a perceptual rating on general pronunciation accuracy. In addition, learners' gesture performance accuracy during the training phase was also rated.

- Study 3 (Chapter 4) trains Catalan learners of French, with an elementary to intermediate proficiency in French, prosodic and vocalic features in a classroom-based setting. It evaluates the effects of visuospatial hand gestures encoding pitch and durational properties at phrase-level on the global pronunciation proficiency, and the pronunciation of front rounded vowels. We predict that with embodied prosodic training, Catalan learners of French would achieve (a) better global pronunciation (assessed by accentedness, comprehensibility, and fluency) and (b) increased accuracy in the production of front rounded vowels

/y, ø, œ/. Fifty-seven Catalan learners of French practiced pronunciation in one of two conditions: one group observed visuospatial hand gestures embodying prosodic features of the sentences that they were listening to, while the other group did not see any such gestures. The learning outcome was assessed in a pretest, a posttest (one week after training), and a delayed posttest (two weeks after training) through a dialogue-reading task and a sentence imitation task. The quality of the front rounded vowels was acoustically assessed with a formant analysis, while the global pronunciation proficiency was perceptually assessed by native speakers for accentedness, comprehensibility, and fluency.

The three independent studies tested the main hypothesis from complementary angles. We explored the effects of visuospatial hand gestures in both segmental and suprasegmental domains with beginners (Studies 1 and 2) and more experienced learners (Study 3). Methodologically, we proposed both laboratory settings (Studies 1 and 2) and a more classroom-based environment (Study 3), and used a variety of training and testing materials involving controlled word or sentence imitation tasks (Studies 1-3), and a less controlled discourse-reading task (Study 3). The assessment of the pronunciation quality involved a combination of perceptual ratings on accentedness, comprehensibility, and fluency; and acoustic measures such as vowel duration (Study 1), Voice Onset Time (VOT) (Study 2), and vowel formants (Study 3). Moreover, the target L2 languages (i.e., Chinese, French, and Japanese) were typologically and phonologically distinct. All in all, we expect that the results of the three

studies can be generalized to second language learning and that our results can contribute to increasing the body of evidence in favor of embodied training in second language pronunciation.

2

CHAPTER 2: DURATIONAL HAND GESTURES FACILITATES THE PRONUNCIATION OF VOWEL-LENGTH CONTRASTS

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

2.1 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) and non-referential 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.

2.1.1 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 ten 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 twelve 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 fourteen-month classroom study, Macedonia and Klimesch (2014) investigated whether using iconic and metaphoric gestures helped L2 lexical learning. They taught 36 non-words 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.

2.1.2 Benefits of beat gestures in L2 vocabulary and pronunciation learning

Beat gestures are a type of non-referential hand gesture which 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. In order 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; Y. Zhang et al., 2020).

2.1.3 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, Wang, Jongman, & Sereno, 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. (2017) 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 which 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.

2.1.4 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 et al., 2014; Hirata & Kelly, 2010; 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 ten-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 abovementioned 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 abovementioned 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 themselves 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 or not. 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 which can be produced by the hands.

2.1.5 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 of 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/, /ɛ/, /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*. Since 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 native-like ratio of long to short vowel durations in Japanese speech production.

2.1.6 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), since musicians are more sensitive to subtle changes in linguistic pitch than non-musicians

(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 non-lexical pitch patterns strongly correlate with the learning of lexical pitch patterns (M. 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-Garcia 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 towards 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.

2.2 Methods

The experiment consisted of a between-subjects training session with a pretest–posttest design, where participants were trained with ten 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.

2.2.1 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.

2.2.2 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.

a) Audiovisual materials for the familiarization phase

For the familiarization phase, a short 1.5-minute audiovisual sequence was created in order 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 ten 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., *íto* '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 perceive (Tajima et al., 2008). All the syllables in the target words complied with the phonotactic constraints of Catalan.

[2] The 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 onwards (in this study, LH(H) pattern). (See Pierrehumbert & Beckman, 1988, pp., 7-8 for more details).

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
joko	*yoko*	side	*joko:*	*yoko:*	rehearsal
ɕaɾe	*xare*	joke	*ɕaɾe:*	*xare:*	reward
kaze	*kaze*	wind	*kaze:*	*kaze:*	taxation
goke	*goke*	widow	*goke:*	*goke:*	word form
toko	*toko*	bed	*toko:*	*toko:*	voyage
íto	*ito*	thread	*íto:*	*ito:*	east
ʤíko	*tgiko*	accident	*ʤíko:*	*tgiko:*	affairs
kúɾo	*kuro*	black	*kúɾo:*	*kuro:*	troubles
kádo	*kado*	corner	*kádo:*	*kado:*	art of poetry
ído	*ido*	water well	*ído:*	*ido:*	medicine

[a] The phonemic transcription conformed to Catalan orthography to facilitate reading by Catalan-speakers.

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 × 2 length contrasts × 2 conditions × 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.

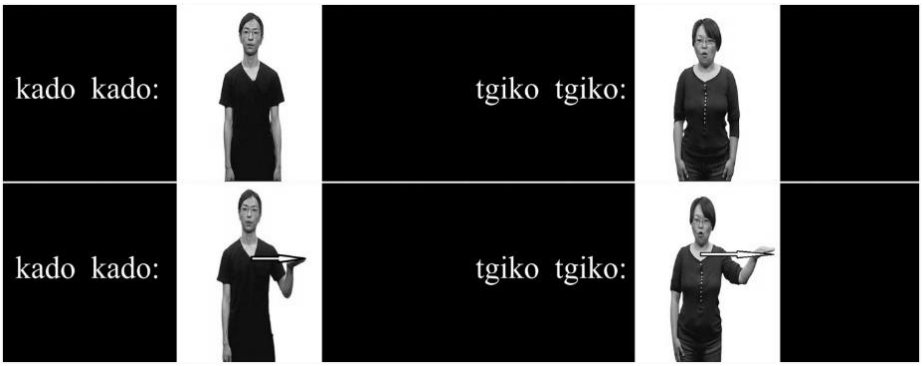
After recording, the videos were edited with Adobe Premiere Pro CC 2018 software. First, the videos were digitally flipped so that the movement appeared to be made with the left hand and participants could mirror the gestures with right hands. In order to control for any potential differences in the 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 five-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 ten pairs of words (block 1) followed by a repetition of these ten 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 ten 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).

Figure 1

Screenshots of Two Trials of the Training Session in NG Condition (Upper Panel) and in G Condition (Lower Panel).



Note. 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.

b) 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.

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
toko/toko: (bed/voyage) joko/joko: (side/rehearsal) kádo/kádo: (corner/poetry art)	[M] Koɾe-ga ___ to jomu. 'This is pronounced as __'	[F] Aɾe-ga ___ deaɾu. 'That is ___'
ído/ído: (water well/medicine) ʤíko/ʤíko: (accident/affairs)	[F] Soko-wa___ ga nai. 'There does not exist ___'	[M] Soko-wa ___ ga aɾu. 'There exists ___'
sotsu/sotsu: (miss/ communication) oɾe/oɾe: (I-masculine/thank)	[M] Aɾe-ga ___ deaɾu. 'That is ___'	[F] Koɾe-ga ___ to jomu. 'This is pronounced as __'
mizo/mizo: (ditch/ unprecedented) kíjo/kíjo: (service/skillful) ɾíka/ɾíka: (science/liquor)	[F] Soko-wa___ ga aɾu. 'There exists ___'	[M] Soko-wa ___ ga nai. 'There does not exist-'

Note. [F] = female speaker; [M] = male 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 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 × 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 of 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.

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

[3] http://www.surveygizmo.com

ɕaɾe/ɕaɾe: (joke/reward) kaze/kaze: (wind/taxation) goke/goke: (widow/word form)	[M] Koɾe-ga ___ deaɾu. 'This is ___.'	[F] Aɾe-ga ___ to jomu. 'That is pronounced as ___.'
íto/íto: (thread/to the east of) kúɾo/kúɾo: (black/troubles) sake/sake: (wine/leftist) iso/iso: (beach/transference) áse/áse: (sweat/Mencius)	[F] Koɾe-ga ___ deaɾu. 'This is ___.'	[M] Aɾe-ga ___ to jomu. 'That is pronounced as ___.'
íɕo/íɕo: (suicide note/clothes) sáju/sáju: (hot water/left-right)	[M] Koɾe-ga ___ deaɾu. 'This is ___.'	[F] Aɾe-ga ___ to jomu. 'That is pronounced as ___.'

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

c) 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 or not 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 non-melodic drum-beats, 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 or not. 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 (Y. Zhang et al., 2020). A total of 24 lists composed of commonly-used Catalan words were selected as the test materials (see Table A1). 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 which were embedded into a PowerPoint presentation.

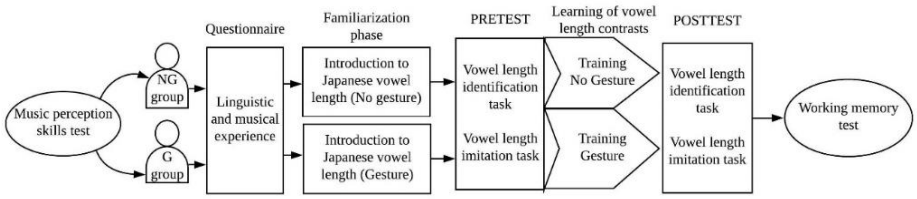
2.2.3 Procedure

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 participants 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.

Figure 2

Experimental Procedure



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 above. 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).

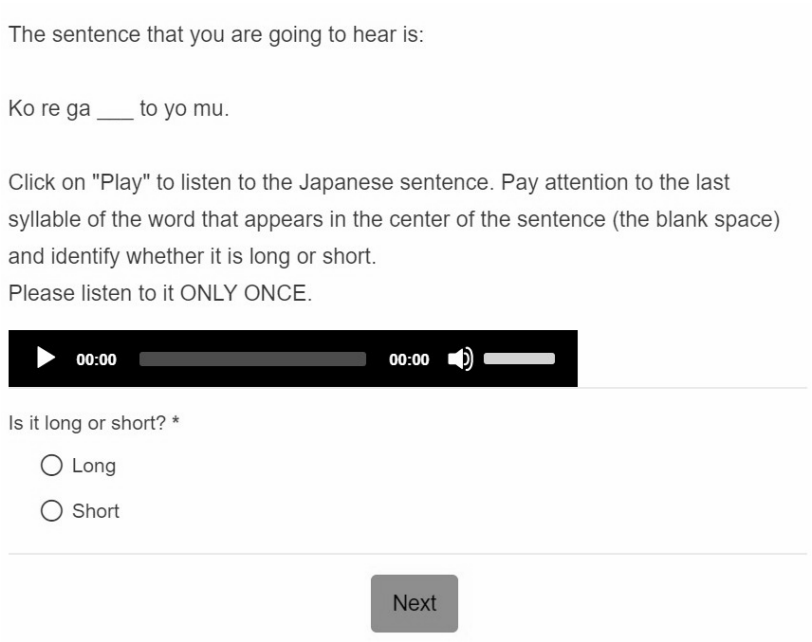
Music perception skill tests. The day before the experiment, participants were sent a link to access the rhythm and pitch tests online. Upon finishing the tests, their scores were automatically generated and exported from PROMS. The full procedure lasted approximately 15 minutes.

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 they were assigned to. 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 twenty audio items were automatically randomized by the software.

Figure 3

Screenshot of a Sample Page from the Vowel-Length Identification Task (English Translation).

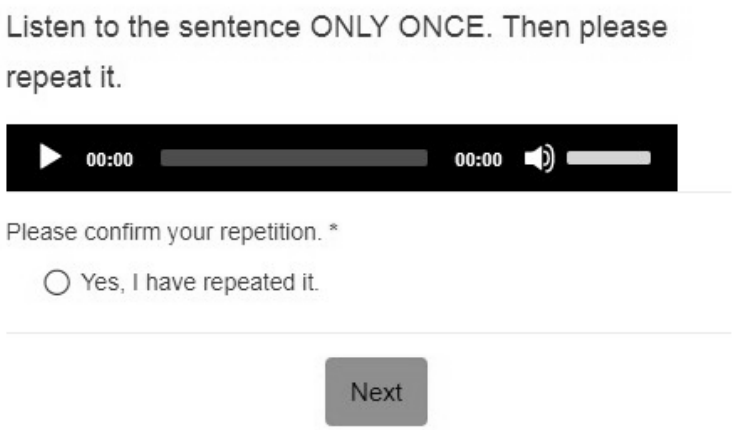


The target words, instructions, and procedure were the same for pretest and posttest. However, as noted above, the order of carrier sentences and speakers varied across tests.

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 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.

Figure 4

Screenshot of a Sample Page from the Vowel-Length Imitation Task (English Translation)



The target words and testing procedure were identical for pre- and post-test, 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.

2.2.4 Coding of the data

a) Vowel-length identification task

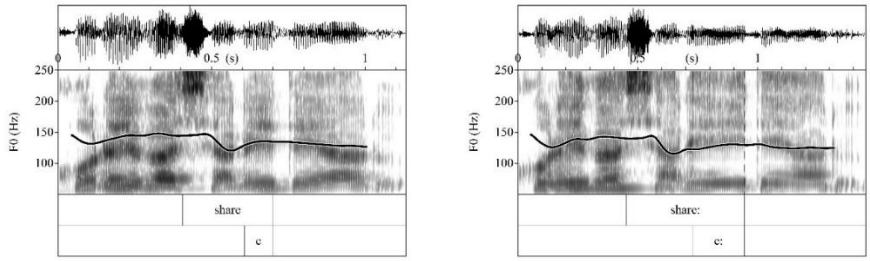
Participants' responses were assessed according to a binary rating system whereby a correct answer was given a score of '1' and an incorrect answer '0'. The 'Accuracy Rate' was obtained by calculating the percentage of correct answers over the total number of trials for each participant, with separate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

b) Vowel-length imitation task

In order to acoustically assess participants' performance on vowel-length contrasts, participants' oral productions during pre- and posttest were analyzed using PRAAT software (Boersma & Weenink, 2017). 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).

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.



Note. The two tiers are the following: target words ('share' and 'share:') and starting and ending points of the target vowels ('e' and 'e:').

After annotation, the duration of each labeled vowel was automatically extracted by means of a PRAAT script.[4] 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.

c) Musical measures

The pre-experimental questionnaire elicited information about each participant's musical background (see Table A2). 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

[4] The script was created by Mietta Lennes and modified by Dan McCloy.

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 five-point scale, ranging from 1 for 'non-musician' to 5 for 'professional musician', which was then labelled as "Self-Perceived Musical Skills".

Regarding musical aptitude, participants' scores on the music perception skill tests of pitch (labelled "Pitch Perceptual Ability") and rhythm (labelled "Rhythm Perceptual Ability") were automatically generated by the PROMS online testing system. In order to generate a categorical variable, a Two-Step 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.

d) 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.

2.2.5 Statistical analysis

The statistical analysis was carried out using IBM SPSS Statistics 24 (IBM Cooperation, 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.

2.3 Results

2.3.1 Vowel-length Identification Task

Table 4

Estimated Mean, Std. Error and 95% Confidence Interval for the Accuracy Rate (%) at Pretest and Posttest Across Conditions

		Estimated		95% Confidence Interval	
Condition	Test	Mean	Std. Error	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.

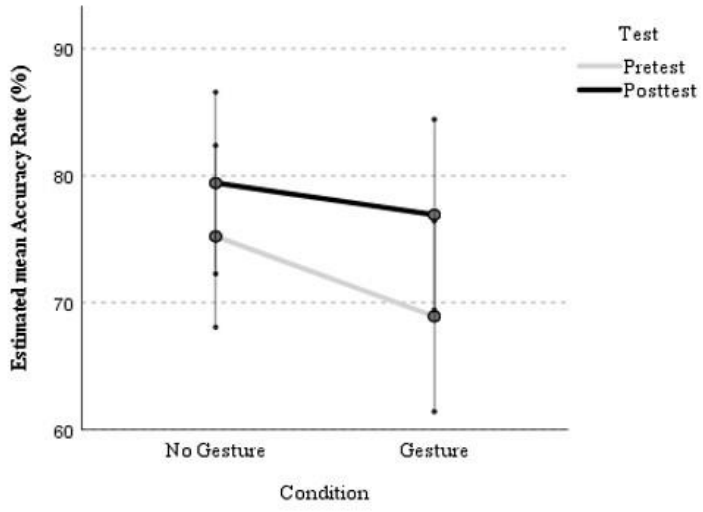


Table 4 and Figure 6 show the mean Accuracy Rate obtained for the vowel-length identification task across conditions (NG and G) and tests (pretest and posttest). The descriptive data show that participants in the G group improved more (*Contrast estimate* = 8.000%) than those in the NG group (*Contrast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

Table 5

Summary of GLMM: Fixed Effects for the Mean Accuracy Rate of Identification Task

Fixed factors	*F*	*df1*	*df2*	*p.*
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

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 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.

2.3.2 Imitation Task

Table 6

Estimated Mean, Std. Error and 95% Confidence Interval for the Duration Ratio at Pretest and Posttest Across Conditions

		Estimated		95% Confidence Interval	
Condition	Test	Mean	Std. Error	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

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.

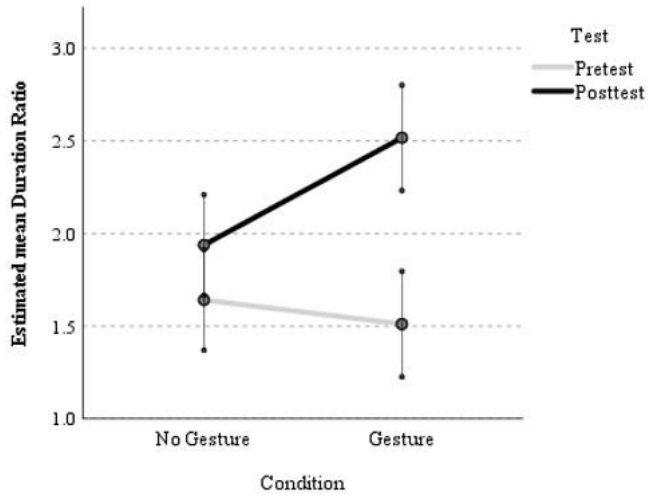


Table 6 and Figure 7 show the Mean Duration Ratio from the vowel-length imitation task across conditions (NG and G) and tests (pretest and posttest). The improvement in the Mean Duration Ratio from pretest to posttest for the G group (*Contrast estimate* = 1.006) was larger than that for the NG group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

Summary of GLMM: Fixed Effects for the Mean Duration Ratio of Imitation Task

Fixed factors	*F*	*df1*	*df2*	*p.*
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

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 × 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, i.e., 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.

2.4 Discussion and Conclusion

The present study examined the effectiveness of visuospatial hand gestures depicting vowel-length features on perceiving and producing non-native 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 et al., 2014; Hirata & Kelly, 2010; 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 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 non-natives 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 dis-

tractions across the two tasks. As noted above, in order 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. Since 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. On the other hand, since 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 (Paula Roncaglia-Denissen et al., 2016). 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 which 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 non-words 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 non-native 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 which 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.

3

CHAPTER 3: TRAINING NON-NATIVE ASPIRATED PLOSIVES WITH HAND GESTURES: LEARNERS' GESTURE PERFORMANCE MATTERS

Li, P., Xi, X., Baills, F., & Prieto, P. (2021, in press). Training non-native aspirated plosives with hand gestures: Learners' gesture performance matters. *Language Cognition and Neuroscience*. 10.1080/23273798.2021.1937663

3.1 Introduction

The effects of multimodal training involving gestural input in second language (L2) acquisition have become an essential line of research (Gullberg, 2014). Given that gestures are intensively used in L2 classrooms (Hudson, 2011; Smotrova, 2017), more empirical evidence is needed to assess whether they are effective for L2 pronunciation and vocabulary learning. In this training study, we explore the role of performing visuospatial gestures that mimic phonetic features in L2 pronunciation and vocabulary learning by training Catalan speakers to learn Mandarin aspirated plosives and Mandarin words containing these phonemes. In what follows, we summarize a series of theories that may explain the effects of multimodal learning involving gestures and the role of gestures in L2 learning, focusing on pronunciation and vocabulary.3.1.1 Benefits of Gestures in L2 Vocabulary Learning

3.1.1 Theoretical Background

Several theoretical frameworks support the beneficial role of multimodality in second language pronunciation training. The Dual Coding theory supports the role of visual cues. According to this theory, people process verbal and visual information via different but interdependent channels, leading to better learning outcomes since across-modal cues and redundant information can reinforce the learning (Clark & Paivio, 1991; Paivio, 1991). Empirical work has revealed positive evidence for the role of visual cues in L2 pronunciation learning (e.g., Hardison, 2004; Hazan et al., 2005; Olson, 2014). The role of gesture is further supported by the Embodied/Grounded Cognition theory, which holds that body and mind are two integrated systems involved in the human cognitive process (Barsalou, 2008; Ionescu & Vasc, 2014). There is evidence that language

is embodied (Desai et al., 2010; J. Yang & Shu, 2016), while gestures, which are closely tied to speech (Iverson & Goldin-Meadow, 2005; McNeill, 1992), may stem from spatial representations and mental images and may arise from an embodied cognitive system (Hostetter & Alibali, 2008). Furthermore, according to the Cognitive Load theory, a proper instructional design should minimize learners' cognitive load, allowing them to make more efforts to process the learning materials (Paas & Sweller, 2012; J. Sweller et al., 1998). As people usually shift information to the body or the environment to reduce cognitive load (Risko & Gilbert, 2016), body movements, such as hand gestures, can lighten the cognitive load and save cognitive resources for learners to improve their learning performance.

Taken together, these theories suggest that active use of body movements, especially hand gestures, should be encouraged in teaching practice, as gestures function as visual cues as well as a manifestation of embodied language. However, empirical evidence is still needed to evaluate the implications of embodied cognition in the training of L2 pronunciation. Therefore, exploring the role of gestures in L2 pronunciation training may provide a direct test on the predictions of embodied cognition in L2 phonology.

3.1.2 Effects of Visuospatial Hand Gestures in L2 Perception and Production

A series of training studies have shown that visuospatial hand gestures (i.e., hand movements that represent suprasegmental and/or segmental features of a language in space) may affect the perception and production in an L2 in various aspects.

a) Visuospatial hand gestures in L2 perception

First, pitch gestures (e.g., hand gestures depicting the F0 contour in space) were found to improve the perception of L2 lexical tones. For example, Baills et al. (2019) demonstrated that observing and producing pitch gestures favored the learning of L2 Chinese lexical tones at the perceptual level and the acquisition of word meanings (see also Morett & Chang, 2015 for similar results). Zhen et al. (2019) confirmed that when the pitch gestures were performed horizontally other than vertically in space, producing hand gestures was more helpful than merely observing them. This finding points to the importance of gesture form and performance during training. Moreover, observing pitch gestures has also been found to boost the perception of L2 intonation (Kelly et al., 2017).

By contrast, a handful of studies have claimed that gestures illustrating durational features were not helpful in the perception of L2 vowel-length contrasts (Hirata et al., 2014; Hirata & Kelly, 2010; Kelly et al., 2017; P. Li, Baills, et al., 2020). At the segmental level, gestures mimicking specific phonetic features also do not seem helpful in the perception of target phonemes, such as aspirated consonants (Xi et al., 2020).

b) Visuospatial hand gestures in L2 production

Mixed results have also been found regarding the role of visuospatial hand gestures in L2 speech production.

First, observing rhythmic beat gestures may reduce learners' foreign accents (Gluhareva & Prieto, 2017), and producing them may help L2 pronunciation more than merely observing them (Kushch, 2018). Second, observing pitch gestures seems to favor the production of L2 intonational

features (Yuan et al., 2019). As for durational features, producing horizontal hand sweep gestures has been suggested to improve the production of long and short vowels (Li et al., 2020). However, beat gestures performed on the stressed syllables failed to improve the lexical stress production in an L2 (van Maastricht et al., 2019), and gestures mimicking pitch contours of Chinese lexical tonal patterns were also not helpful in simultaneous speech production (Zheng et al., 2018).

Turning to the production of L2 segmental features, to our knowledge, only three experimental studies have tested the effects of observing visuospatial hand gestures cueing phonetic features on L2 pronunciation learning. First, Amand and Touhami (2016) found that observing gestures could help French speakers pronounce English unreleased plosives. The gestures for the released plosives were a fist-to-open hand gesture and, for the unreleased ones, a stretched-fingers-to-fist gesture. More recently, Hoetjes and van Maastricht (2020) compared the effects of observing pointing gestures and gestures mimicking articulatory information on the learning of two Spanish segments, /u/ and /θ/, by Dutch speakers. The results revealed that pointing gestures had a positive effect on the pronunciation of both /u/ and /θ/, and that gestures conveying articulatory information facilitated the pronunciation of /u/ but hindered the pronunciation of /θ/, suggesting the importance of gesture type on L2 segmental learning. Xi et al., (2020) trained 50 Catalan speakers to learn six pairs of Mandarin consonants with or without gestures. Three pairs were plosives /p-pʰ, t-tʰ, k-kʰ/ which contrast in aspiration (i.e., in the absence or presence of a strong air burst) whereas the other three pairs were the affricates /ts-tsʰ, tɕ-tɕʰ, tʂ-tʂʰ/, which are phonologically described as unaspirated-aspirated contrasts but differ acoustically in the duration of frication as well. A fist-to-open-hand gesture was used to

simulate the extra airburst of the aspirated plosives. However, this gesture was deemed an inadequate visual representation of the longer duration frication of the aspirated affricates. The results revealed that while observing this gesture significantly improved participants' pronunciation of aspirated plosives, it failed to help the pronunciation of affricates. This not only constituted new evidence supporting the positive role of hand gestures in L2 pronunciation learning but again suggested that the form of a gesture must be appropriate to the specific phonetic features it is intended to represent.

In short, it seems that, despite some mixed results, observing and producing hand gestures benefits the learning of a variety of L2 suprasegmental features, but further empirical evidence is needed to assess the role of hand gestures mimicking phonetic features, especially segmental features. First, it is not clear whether producing hand gestures benefits segmental learning since none of the abovementioned studies (Amand & Touhami, 2016; Hoetjes & van Maastricht, 2020; Xi et al., 2020) asked learners to perform gestures during training (as opposed to merely observing them) although this technique has been shown to be effective in multimodal learning contexts (Macedonia, 2019). Second, there is a lack of information about how the accuracy of self-performed gestures may impact pronunciation learning. Third, few studies have assessed whether the effects of visuospatial hand gestures on L2 segments are maintained over time.

3.1.3 Effects of Hand Gestures on L2 Vocabulary Learning

Observing representational gestures that depict the referent (e.g., iconic and metaphorical gestures) has been shown to facilitate L2 vocabulary

learning. In her pioneering study, Allen (1995) found that gestural training could help learners memorize L2 idiomatic expressions better than non-gestural training. Later, Kelly et al. (2009) showed that learners benefited from observing iconic gestures which were congruent with the meaning of L2 words as opposed to incongruent gestures. However, the effects of representational gestures are constrained by phonology: when the phonological demands are high (e.g., minimal-pair words), iconic gestures hindered the memorization of these words; while when the contrast is easy (e.g., Japanese /tate/ vs. /butta/), iconic gestures are helpful (Kelly & Lee, 2012).

Apart from merely observing hand gestures, a series of studies have shed light on the role of gesture production on L2 word memorization (e.g., Macedonia et al., 2011; Macedonia & Klimesch, 2014). Notably, although producing and observing iconic hand gestures has been found to equally benefit L2 word recall (N. Sweller et al., 2020), spontaneous gesture production seems to be more effective than non-spontaneous gesture observation on L2 word memorization (Morett, 2018).

The positive role of hand gestures in L2 vocabulary learning is not limited to representational gestures. Gestures that do not encode semantic meaning can also help enhance memorization. Especially with adults, both iconic and beat gestures were found to benefit the word recall (So et al., 2012). Later, Kushch et al. (2018) confirmed that observing beat gestures illustrating speech prominence helped memorize L2 words. Interestingly, pitch gestures depicting the tonal patterns could also boost memorizing L2 Mandarin words contrasting in lexical tones (Baills et al., 2019; Morett & Chang, 2015).

In short, although gestures encoding semantic or suprasegmental information could help the learning of L2 vocabulary, little research has been done on the effects of visuospatial hand gestures encoding phonetic features in this domain. Furthermore, it remains an open question whether the effects of these gestures on vocabulary learning can be maintained over time since most of the experiments in this field (Baills et al., 2019; Kushch et al., 2018; Morett & Chang, 2015) have only tested for learning effects immediately after training.

3.1.4 Goals of the Present Study

The present study investigates the possible benefit of performing a fist-to-open-hand gesture mimicking the air burst of Mandarin aspirated plosives for learning the pronunciation of these sounds. We selected three pairs of aspirated vs. unaspirated plosives, /p-p^{h}, t-t^{h}, k-k^{h}/ (Duanmu, 2007), to be the target items. Since the participants were Catalan speakers, and Catalan plosives /p-b, t-d, k-g/ do not contrast in aspiration (Wheeler, 2005), we hypothesized that the participants would find it challenging to produce this contrast based on previous findings on European learners learning Chinese (N. F. Chen et al., 2013). The gesture for cueing these aspirated plosives (see Figure 1) was adapted from Xi et al. (2020) and Y. Zhang (2002). No gesture was provided for unaspirated plosives since unaspirated plosives are already part of the Catalan consonant inventory.

Figure 1

The Fist-to-Open-Hand Gesture for Aspirated Plosives

We, therefore, addressed the following two research questions:

RQ1: Does producing visuospatial hand gestures cueing phonetic features favor *L2 segmental learning*? We would compare the effects of training with and without gestures on both perception and production while assessing the accuracy of learners' gesture performance on the one hand and delayed learning effects on the other.

RQ2: Does producing visuospatial hand gestures cueing phonetic features favor the *recognition of novel words* displaying the target phonemes, and are the learning effects maintained over time? Here we would compare the effects of training with and without gestures on word recognition and retention and again see if the accuracy of learners' gesture performance during training had any impact on their learning.

Additionally, two individual factors would need to be controlled for, namely musical experience and working memory, since these two factors have been reported to affect second language learning (Chobert & Besson, 2013; Rota & Reiterer, 2009).

3.2 Methods

The experiment consisted of a between-subjects training session with a pretest–posttest design, where participants were trained with ten 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.

3.2.1 Participants

Sixty-seven undergraduate students (61 females, aged 18–24 years, M_{age} = 19.31 years, *SD* = 1.64) were recruited from a public university in a Catalan-Spanish bilingual area. Each participant reported speaking Catalan at least 75% of the time in daily verbal communication and was thus considered a Catalan-dominant speaker. None of the participants reported hearing impairment. Each of them received €10 in compensation.

Following recruitment, participants were randomly assigned either to one of the two experimental conditions, (a) the No Gesture condition (*n*

= 29, female = 26) and (b) the Gesture condition (*n* = 29, female = 26), or to (c) the Control condition (*n* = 9, female = 9)[5].

The foreign languages that the participants reported speaking included English, French, German, Italian, Portuguese, Romanian, and Russian. No one reported having any prior knowledge of Chinese. Among those foreign languages, only English and German are said to have plosives involving aspiration contrasts, although the contrasts in those two languages have different phonetic realizations (e.g., Chao & Chen, 2008; Kleber, 2018). Note that each group involved a similar portion of participants who spoke each language (see Table 1). Therefore, although our participants represented a multilingual population, the multilingual profile would not seem to cause group differences in learning a new language.

Table 1

Foreign Languages Spoken by the Participants (Number and Percentage) in Each Group.

	No Gesture (*n* = 29)	Gesture (*n* = 29)	Control (*n* = 9)
English	29 (100%)	29 (100%)	9 (100%)
French	22 (76%)	17 (59%)	7 (78%)
German	19 (66%)	21 (72%)	5 (56%)
Italian	22 (76%)	23 (79%)	6 (67%)
Portuguese	4 (14%)	0 (0%)	1 (11%)

[5] The reason for including a control group was to assess whether the changes between tests were due to mere repetition of the testing materials. We believe that a relatively small sample size in the non-training group would not affect the results (see Mora & Levkina, 2018; Saito & Lyster, 2012 for similar design).

Romanian	1 (3%)	2 (7%)	0 (0%)
Russian	1 (3%)	3 (10%)	0 (0%)

3.2.2 Materials

In this section, we describe the creation of the materials used in the experiment. All the audio recordings were performed in a radio studio with professional equipment and later edited with Audacity 2.1.2, while all the audio-visual materials were prepared in a professional video-recording studio using a PDM660 Marantz professional portable digital video recorder and a Rode NTG2 condenser microphone and later edited with Adobe Premiere Pro CC 2018. After preparation, all the materials were uploaded to SurveyGizmo (https://www.surveygizmo.com), an online platform to create the training and testing webpages.

a) Audio-visual materials for the familiarization phase

For this phase, separate videos were created for each of the three conditions (Gesture, No Gesture, and Control). In all three versions, a Chinese language instructor introduced the three pairs of Mandarin plosives contrasting in aspiration and then instructed participants how they should perform in the training and the tasks.

b) Audio-visual stimuli for the pronunciation training session

The stimuli for the pronunciation training session were six pairs of Mandarin disyllabic words contrasting only in consonantal aspiration, which was located in word-initial position (see Table B1).

Two right-handed native Mandarin instructors (one female) were video-recorded producing the target words. They were filmed against a white

background, with the upper half of their body and face visible so that lip and mouth articulatory movements would be clearly seen. For the No Gesture condition, the instructors were asked to produce the words in a natural way and without any body movements other than those strictly related to oral articulation. For the Gesture condition, they were asked to produce a fist-to-open-hand gesture to visually mimic the burst of air as they uttered the aspirated plosives (see Figure 1) while keeping the rest of the body still. They were asked to first raise their two hands to the height of the shoulders, and once they had reached this height, to open their palms quickly towards the camera. Crucially, they were asked to use both hands to make the visual cue more salient and also to avoid possible interferences due to hand preference.

A total of 36 video clips were obtained (6 words with unaspirated plosives × 2 instructors + 6 words with aspirated plosives × 2 instructors × 2 conditions). In order to avoid any potential differences in speech across the two conditions, the audio track of the No Gesture videos was copied onto the corresponding audio track of the Gesture videos. To check whether the audio track and the image of the video clips were temporally synchronous, three native speakers of Mandarin evaluated the video clips using a 5-point Likert scale (from 1 "Not synchronous at all" to 5 "Very synchronous"). The resulting mean rating was very high (M = 4.72, SD = 0.51).

Finally, the 36 video clips were used to create two training videos, one for each condition. In both videos, the clips were organized into three sequences. The first sequence was designed to train participants to repeat the words in isolation. For each word, the Catalan transcription of the word first appeared on the screen (2 s), then an instructor uttered the

word (3 s), followed by a black screen displaying “Repeat that” (2 s). Next, the other instructor uttered the same word again (3 s), which ended with a black screen saying “Repeat that” (2 s).

By contrast, the second and third sequences were designed to train the words in pairs, with each pair of words trained once in each sequence. Each trial began with the Catalan transcription of the word pair (2 s). Then, the two instructors appeared in turn uttering each of the words (6 s). The trial ended with a black screen displaying “Repeat that” (4 s).

c) Auditory stimuli for the identification and imitation tasks

Six pairs of Mandarin words featuring the aspiration contrast in word-initial position were selected for the identification task. Half of the words were included in the pronunciation training phase, and the other half were not (see Table B2).

The stimuli for the imitation task also consisted of six pairs of Mandarin words, three pairs being trained while the other three, untrained (see Table B3).

The same two instructors who produced the training videos recorded the items for the two testing tasks. For each task, half of the word pairs were spoken by the male speaker and the other half by the female speaker.

d) Audio-visual stimuli for the vocabulary training session

Six pairs of monosyllabic Mandarin words expressing common everyday meanings were selected for the vocabulary training session (see Table B4). Each word pair contrasted only in aspiration of the word-initial plosive consonants.

The creation of the video clips for the vocabulary training session followed the same procedure as that of the pronunciation training. Similar to the materials of the pronunciation training, the mean rating of the temporal synchrony was 4.69 (*SD* = 0.52) and was thus considered very high.

Each trial was sequentially organized as follows: First, the Catalan translation of one target word appeared on the screen (3 s), followed by one instructor uttering the training word (2 s) and ended with a short instruction saying "Repeat that" (3 s). Each of the training words was trained three times, which was embedded in three video sequences with different orders.

e) Auditory stimuli for the word-meaning association task

The stimuli for the word-meaning association task were the same 12 training words as those used in the vocabulary training session. The materials were prepared following the same procedure as the identification and imitation tasks.

f) Control tasks

First, participants' musical experience was assessed by means of a questionnaire adapted from Boll-Avetisyan et al. (2017) and Li et al. (2020) (see Table B5). Second, participants' working memory was measured by means of a classic digit span task (Wen, 2018). To keep the duration of the experiment reasonable, only a forward digit span task was chosen. Following Woods et al., (2011), the task was embedded in a program developed using PsychoPy3 software (Peirce et al., 2019). Additionally, participants were asked to evaluate their motivation for learning Chinese on a 9-point Likert scale (1 = "not at all", 9 = "very much").

3.2.3 Procedure

Participants finished the learning procedure individually in an experimental room with a laptop computer. Prior to beginning the experiment, all participants signed a consent form which allowed the researchers to process their personal data and gave their permission to be video-recorded during the whole procedure using Camera software. The video recording was done in order to allow the researchers to gather data on participants' pronunciation and gesture performance.

To begin the experiment, participants first completed the questionnaire about their linguistic background and musical experience (5 min) then viewed the familiarization video (2 min). This was followed by the pretest tasks (about 6 min). The pretest involved an identification task and an imitation task. In the identification task, participants listened to each of the 12 words only once and had to identify whether the target word started with an aspirated or an unaspirated sound by choosing from two options written in Catalan transcription (e.g., *kuli* vs. *kʰuli*, see Table B2). Then, in the imitation task, participants listened to each of the 12 Mandarin words once and imitated each of them right after the model speech. In both tasks, the items were presented in random order.

Next, participants in the two experimental conditions watched the pronunciation training video (5 min) and either repeated the Mandarin word pairs aloud in the No Gesture condition or repeated them aloud while performing gestures in the Gesture condition. By contrast, those in the control condition watched a 5-minute video of a symphony orchestra playing instrumental music (Sabkay71, 2011).

Following this session, all participants completed the immediate posttest (around 6 min), which was exactly the same as the pretest.

Then, the two experimental groups were exposed to the vocabulary training session (5 min). Participants were asked to repeat the words (with or without gestures) and memorize their meaning. The training session was immediately followed by a word-meaning association task (around 3 min), where participants listened once to each of the 12 training words and had to choose the correct Catalan translation from three options: The correct translation for the testing word, the translation of the testing word's counterpart which contrasts with it in aspiration, and the translation for another training word. However, the control group was not involved in either the vocabulary training session or the vocabulary test.

At the end of the experiment, all participants took the forward digit span test by recalling a length-increasing (3-16) sequence of digits in a forward order (around 5 min, see Woods et al., 2011 for details).

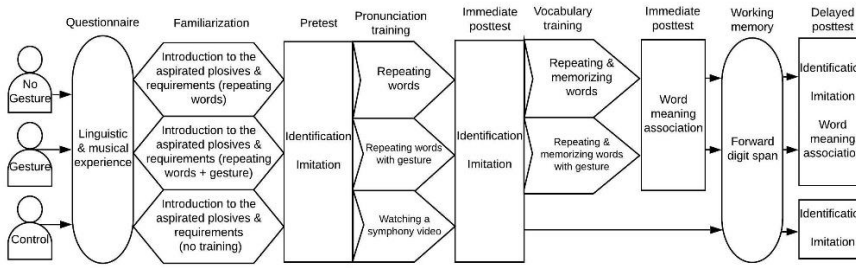
Three days later, a delayed posttest was administered (around 10 min), in which all the participants repeated the identification and imitation tasks while the participants in the No gesture and Gesture conditions also repeated the word-meaning association task.

Overall, the duration of the whole experiment, including the delayed posttest, was about one hour. Except for the forward digit span test, the rest of the experiment was done via SurveyGizmo. Participants were allowed to set the volume at their most comfortable level and self-paced the whole learning procedure online. No feedback was provided during the entirety of the experiment.

A schematic diagram of the experimental procedure can be seen in Figure 2.

Figure 2

Experimental Procedure



3.2.4 Data Coding

a) Identification task

The task was assessed using a binary rating system: a correct answer was marked as 1 and an incorrect answer, 0. The responses of the participants were exported from SurveyGizmo and then labelled as "identification score".

b) Imitation task

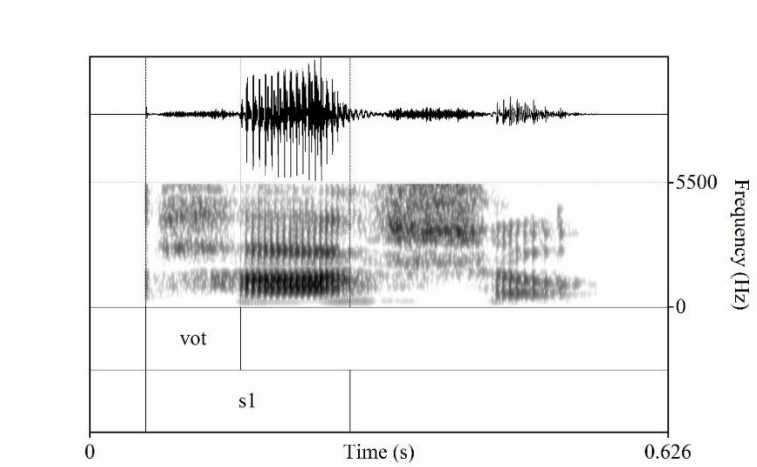
A total of 2,412 recordings were obtained from the imitation task. The recordings were acoustically analysed for the plosives and also perceptually rated for overall pronunciation.

Acoustic Analysis. In acoustic phonetics, Voice Onset Time (VOT) is often employed to describe the delay of voicing onset following the released burst, and aspirated plosives have a longer VOT period than unaspirated plosives (Johnson, 2011). Therefore, the first author labelled

the VOT of all the 2,412 initial consonants and the first syllable of each word produced by the participants in the imitation task, using Praat software (Boersma & Weenink, 2020). The VOT was labelled from the onset of the release burst (the release bar in the spectrogram in Figure 3) to the onset of the vocal fold vibration (the first vertical striation in the spectrogram in Figure 3). After annotation, the VOT value and the duration of the first syllable of each word were extracted from Praat. To normalize the data for speech rate, the VOT of each consonant was divided by the first syllable duration (see Boucher, 2002; Whitfield et al., 2018, among many others), yielding a measure of VOT ratio that was used as the dependent variable in our statistical analyses.

Figure 3

Waveform and Spectrogram of a Sample Word páshǒu 'thief'



Note. The "vot" labels the VOT of the target consonant; "s1" labels the syllable containing that consonant.

Perceptual ratings. Five native Mandarin speakers (4 females, aged 21–30 years, M_{age} = 25.80 years) rated the 2,412 words produced by the participants. Before rating, all raters were trained in a 30-minute session to

familiarize them with the evaluation system. Raters were asked to listen to each word and evaluate the pronunciation of the word on a scale from 1 (Not accurate at all) to 9 (Definitely accurate). All the recordings were presented to the raters randomly. Inter-rater reliability was checked by Cronbach's alpha. The results revealed good agreement across the five raters (α =.89). The raters' ratings of each item were then averaged, yielding a pronunciation score.

c) Word-meaning association task

Participants' answers in the word-meaning association task were coded as follows: (a) 1 point was awarded for recognizing the correct translation of the target word, (b) 0.5 point was awarded for recognizing the counterpart of the target word, and (c) 0 point was awarded if the word chosen was not part of the minimal pair. After coding, this score constituted the word recognition score.

d) Gesture performance ratings

Following a thirty-minute training session, three researchers in phonetics and gesture studies (2 females, aged 24–31 years, M_{age} = 27.6 years) rated participants' gesture performance during the pronunciation and vocabulary training sessions, on a scale of 1 (Very bad) to 9 (Very good). Two main evaluative criteria were used, namely (a) the degree of synchrony between the gesture movements and the target aspirated plosives and (b) the degree of similarity between the target gesture shape as performed by the participant and as performed by the instructors. If for a given gesture both of the criteria were fulfilled, the rating was at least 7; if one of the criteria was not satisfied, then the score was 4–6; and if it failed to satisfy both of the criteria, the rating was 3 or less.

A total of 696 video clips from the pronunciation training session and 522 video clips from the vocabulary training session were rated. All the stimuli were presented in a randomized order, and the raters rated each of the items individually. Inter-rater reliability was excellent (α =.93) in the rating of gestures performed during the pronunciation training and good (α =.80) in that of the vocabulary training. Given that the gesture scores were quite dispersed on a 9-point Likert scale, in order to better capture the relationship between gesture performance, test, and aspiration, the score was clustered by a Two-Step Cluster analysis, which is a combination of the two most commonly used cluster methods (Hierarchical and K-means) in L2 research (see Crowther et al., 2021 for a recent synthesis, where 14.5% of the cases directly used Two-Step Clusters). This approach was also adopted by some recent studies on similar topics (Melnik-Leroy et al., 2021; Yuan et al., 2019).

Participants were automatically classified into two different levels according to their gesture scores: (a) Well Performed Gesture group (in the pronunciation training: n = 14, M = 7.07, SD = 0.53; in the vocabulary training: n = 12, M = 7.13, SD = 0.44) and (b) Poorly Performed Gesture group (in the pronunciation training: n = 15, M = 5.16, SD = 0.87; in the vocabulary training: n = 17, M = 5.27, SD = 0.79). A new independent variable was added to the databases, namely "Gesture performance" with four levels: No Gesture, Well Performed Gesture, Poorly Performed Gesture, and Control.

e) Control measures

Participants' answers to the musical experience questionnaire were coded following Boll-Avetisyan et al. (2017) and Li et al. (2020) and

labelled as “musical experience score”. The digit span score (rages 3-16) for each participant was automatically generated by PsychoPy3 software (see Woods et al., 2011 for the scoring system). Finally, the self-estimated motivation was labelled as “motivation score” for further analysis.

3.2.5 Statistical Analysis

Four Mixed-Effects Models were applied to the following outcome measures using the *lme4* package, version 1.1.23 (Bates et al., 2015) in R, version 4.0.2: (a) identification score; (b) VOT ratio; (c) pronunciation score; and (d) word recognition score. The VOT ratio and the pronunciation score were automatically transformed to adjust the normality using the *orderNorm()* function from the *bestNormalize* package version 1.6.1 (Peterson & Cavanaugh, 2019). However, all the descriptive data reported in the Results section were on their original scales.

For all four models, the fixed factors were Gesture performance (four levels: No Gesture, Poorly Performed Gesture, Well Performed Gesture, and Control), Test (three levels: pretest, immediate posttest, and delayed posttest), Aspiration (two levels: aspirated and unaspirated), and their interactions. In order to check that no response biases were affecting the results, we added aspiration (aspirated vs. unaspirated) as a fixed effect in the analyses of identification, pronunciation, and recognition scores. If the participants tended to perform better in one type of item (say the unaspirated), we should expect a significant main effect of aspiration.

The models that best fitted our data were determined by the function *compare performance* from the *performance* package, version 0.4.8 (Lüdecke et al., 2019). For the identification score, the best fitting model

was a Generalized Linear Mixed Model, involving two random intercepts for Participant and for Item. For VOT and the pronunciation score, the best fitting models were two identical Linear Mixed Models, which included a random intercept for Item and a by Participant random slope of Aspiration. For the word recognition score, however, it was a Linear Mixed Model, including two random intercepts for Participant and for Item.

In an initial analysis, we checked if testing items that were familiar to the participants (trained vs. untrained items) had an interaction with the other main effects, three models involving a four-way interaction of Familiarity × Aspiration × Test × Gesture performance were built for (a) identification score, (b) VOT ratio, and (c) pronunciation score. However, no significant four-way interaction, nor significant three-way interaction of Familiarity × Gesture performance × Test was found from any of the measures (all $p > .05$, see Table B6 for the summary of the models). Therefore, we excluded familiarity from the models so that the analyses were nested to our research questions.

In all the models, significance was determined by the Type II Wald chi-squared tests using the *car* package, version 3.0.9 (Fox & Weisberg, 2019). The post-hoc pairwise comparisons were performed with the *emmeans* package, version 1.4.8 (Lenth et al., 2020). The significance of all the contrasts was adjusted for multiple comparisons using the false discovery rate method. The Cohen's *d* was included to assess the effect size (small: $d \geq 0.2$; medium: $d \geq 0.5$; large: $d \geq 0.8$, see Cohen, 1988).

3.3 Results

Because we aimed to determine the effects of gesture performance on the dependent variables, which may vary across tests and aspiration, we mainly focused on the three-way interaction of Aspiration × Test × Gesture performance. If, however, the three-way interaction was absent, we would check the two-way interaction of Test × Gesture performance since whether or not the effects of gesture performance on the dependent variables varied across Test, regardless of Aspiration, was also of interest to us.

3.3.1 Homogeneity among gesture performance conditions in pronunciation and vocabulary training

We ran a series of linear models with gesture performance as the main effect and normalized age, digit span score, and musical score as dependent variables (the normalization was also performed by the *orderNorm()* function). None of the models revealed a significant main effect of gesture performance (all $p > .05$).

In addition, the self-reported motivation score was also analysed in the same way. No significant main effect of gesture performance was revealed, suggesting that participants' motivation of learning Mandarin was similar across groups. Thus, motivation may not be relevant to the performance during training and the learning outcomes at the group level.

Tables B7 and B8 summarize all the descriptive data and statistical results.

3.3.2 RQ 1: Does producing visuospatial hand gestures help the learning of L2 segmental features?

a) The perception of L2 segmental features

Table 2

Means and Standard Deviations of Identification Score Across Test and Group

Condition	Pretest	Immediate posttest	Delayed posttest
	M (SD)	*M (SD)*	*M (SD)*
NG	0.84 (0.37)	0.86 (0.35)	0.85 (0.35)
PPG	0.82 (0.38)	0.84 (0.36)	0.84 (0.36)
WPG	0.82 (0.39)	0.93 (0.26)	0.90 (0.30)
C	0.89 (0.32)	0.88 (0.33)	0.87 (0.34)

Note. NG = No Gesture; PPG = Poorly Performed Gesture; WPG = Well performed Gesture; C = Control.

Table 2 displays the descriptive data of the identification score. There was only a main effect of test, $\chi^2(2) = 6.37$, $p = .041$. Post-hoc comparisons revealed a significant improvement from pretest to immediate posttest, $z = 2.51$, $d = 0.44$, $p = .036$, but no significant contrasts were found between other test pairs. Noteworthily, aspiration was not a significant effect, $\chi^2(1) = 1.08$, $p = .299$, which indicates that participants' responses were not biased towards either aspirated or unaspirated consonants. Moreover, the nonsignificant interactions of Test × Gesture performance, $\chi^2(6) = 9.82$, $p = .133$, and Aspiration × Test × Gesture performance, $\chi^2(6) = 2.56$, $p = .862$, indicate that participants' identification accuracy did not differ between aspirated and unaspirated items across the three tests, regardless of gesture performance.

b) The production of L2 segmental features

VOT analysis. A significant three-way interaction of Aspiration × Test × Gesture performance was found, $\chi^2(6) = 17.40$, $p = .008$. The post-hoc

pairwise comparisons did not show significant contrasts for unaspirated plosives, indicating that the mean VOT of unaspirated plosives remained stable regardless of condition or test. However, for the aspirated plosives, the mean VOT ratio varied across the three tests and the four Gesture performance groups. Specifically, (a) in the No Gesture condition the mean VOT ratio increased significantly from pretest to immediate posttest, $t(2252) = 4.61$, $MSE = 0.07$, $d = 0.49$ $p < .001$, but decreased significantly from immediate posttest to delayed posttest, $t(2252) = -2.59$, $MSE = 0.07$, $d = -0.28$, $p = .015$， although the delayed posttest still showed a significantly higher mean VOT ratio than pretest, $t(2252) = 2.02$, $MSE = 0.07$, $d = 0.22$, $p = .043$. However, (b) in the Poorly Performed Gesture condition, no significant change was found. Contrastingly, (c) in the Well Performed Gesture condition, the mean VOT ratio increased from pretest to immediate posttest, $t(2252) = 4.50$, $MSE = 0.11$, $d = 0.69$, $p < .001$, and from pretest to delayed posttest, $t(2252) = 4.45$, $MSE = 0.11$, $d = 0.69$, $p < .001$. In addition, (d) no significant contrast was found in the Control condition. Table 2 displays the descriptive data of the mean VOT values. Figure 4 shows the mean VOT of the aspirated plosives /p^h, t^h, k^h/ obtained in the imitation task across Gesture performance and Test.

Table 3

Means and Standard Deviations of VOT Ratio of Unaspirated and Aspirated Plosives Across Group and Test

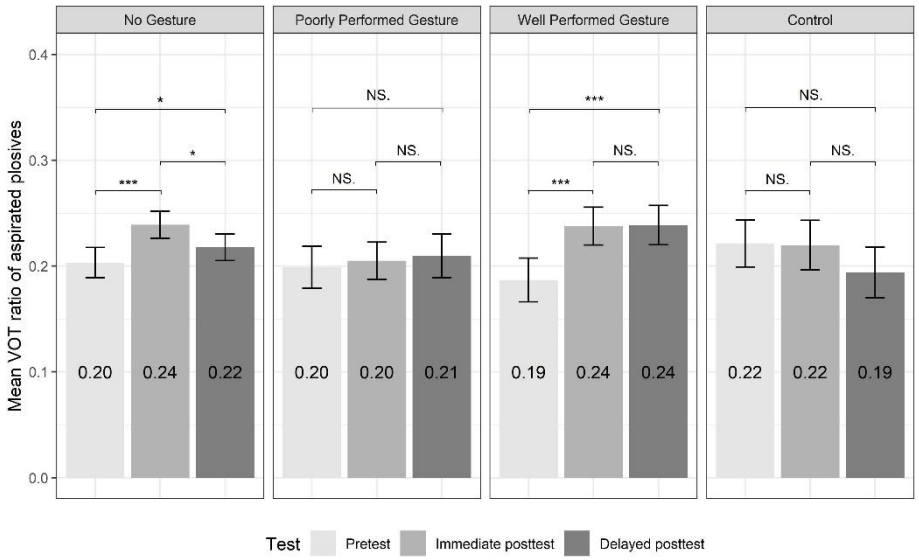
Aspiration	Condition	Pretest	Immediate posttest	Delayed posttest
		M (SD)	*M (SD)*	*M (SD)*
Unaspirated	NG	0.11 (0.10)	0.11 (0.10)	0.11 (0.10)
	PPG	0.10 (0.10)	0.09 (0.10)	0.09 (0.08)
	WPG	0.11 (0.07)	0.10 (0.07)	0.09 (0.08)
	C	0.10 (0.12)	0.11 (0.10)	0.10 (0.10)
Aspirated	NG	0.20 (0.10)	0.24 (0.09)	0.22 (0.08)

PPG	0.20 (0.09)	0.20 (0.09)	0.21 (0.10)
WPG	0.19 (0.10)	0.24 (0.08)	0.24 (0.09)
C	0.22 (0.08)	0.22 (0.09)	0.19 (0.09)

Note. NG = No Gesture; PPG = Poorly Performed Gesture; WPG = Well performed Gesture; C = Control.

Figure 4

Mean VOT Ratio of the Aspirated Plosives Produced by the Participants



Note. The numbers labelled on the bars represent the mean score. Error bars indicate 95% confidence interval. $^{***}p < .001$; $^{**}p < .01$; $^{*}p < .05$; NS. = not significant.

Pronunciation accuracy. The analysis revealed a significant two-way interaction of Test × Gesture performance, $\chi^2(6) = 18.21$, $p = .006$. Post-hoc comparisons showed that in the No Gesture condition, there was a significant improvement from pretest to immediate posttest, $t(2252) = 4.39$, $MSE = 0.06$, $d = 0.33$, $p < .001$, and from pretest to delayed posttest, $t(2252) = 2.89$, $MSE = 0.06$, $d = 0.22$, $p = .006$. In the Poorly Performed Gesture condition the improvement could only be observed from pretest to immediate posttest, $t(2252) = 3.44$, $MSE = 0.08$, $d = 0.36$, $p = .002$. By

contrast, in the Well Performed Gesture condition the mean pronunciation score improved across the three tests, with the immediate posttest outperforming the pretest, $t(2252) = 2.81$, $MSE = 0.09$, $d = 0.31$, $p = .007$, and the delayed posttest outperforming the immediate posttest, $t(2252) = 2.58$, $MSE = 0.09$, $d = 0.28$, $p = .010$, as well as the pretest, $t(2252) = 5.39$, $MSE = 0.09$, $d = 0.59$, $p < .001$. No significant differences across tests were found in the Control condition. Table 3 shows the descriptive data of the pronunciation score. Figure 5 illustrates the mean pronunciation score of the target words across gesture performance and test.

Table 4

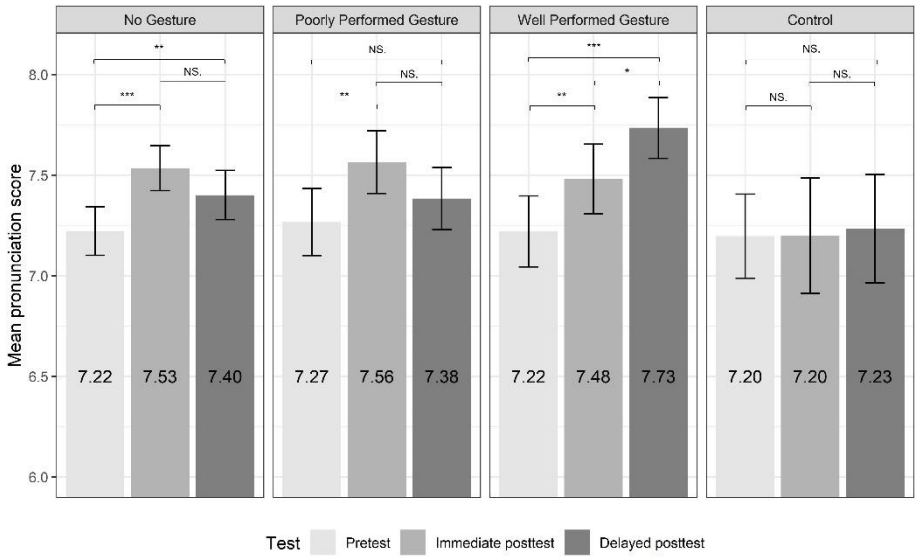
Means and Standard Deviations of Pronunciation Score Across Test and Group

Condition	Pretest	Immediate posttest	Delayed posttest
	M (SD)	*M (SD)*	*M (SD)*
NG	7.22 (1.14)	7.53 (1.06)	7.40 (1.16)
PPG	7.27 (1.13)	7.56 (1.06)	7.38 (1.05)
WPG	7.22 (1.16)	7.48 (1.14)	7.73 (1.00)
C	7.20 (1.10)	7.20 (1.50)	7.23 (1.42)

Note. NG = No Gesture; PPG = Poorly Performed Gesture; WPG = Well performed Gesture; C = Control.

Figure 5

Mean Pronunciation Score of the Target Words Produced by the Participants.



Note. The numbers labelled on the bars represent the mean score. Error bars indicate 95% confidence interval. $^{***}p < .001$; $^{**}p < .01$; $^{*}p < .05$; NS. = not significant.

3.3.3 RQ2: Does producing visuospatial hand gestures help the learning of L2 vocabulary?

The analysis of the word recognition score revealed a significant two-way interaction of Test × Gesture performance, $\chi^2(2) = 7.34$, $p = .025$, indicates differences in word recognition score between conditions across immediate and delayed posttests. Again, aspiration did not reveal a significant main effect, $\chi^2(1) = 0.22$, $p = .636$, which means no response bias was caused by consonantal aspiration. Post-hoc comparisons showed that the No Gesture condition had a significant decay in word recognition scores from immediate posttest to delayed posttest, $t(1260) = -5.12$, $MSE = 0.02$, $d = -0.39$, $p < .001$, whereas the Poorly Performed Gesture condition, $t(1260) = -1.38$, $MSE = 0.03$, $d = -0.14$, $p = .167$, and the Well Performed Gesture condition, $t(1260) = -0.46$, $MSE = 0.04$ $d = -0.05$, $p = .647$, did not. This result indicates that training with gestures helped maintain the recall of the newly learned words. Table 4 shows the

descriptive data of the word recognition score. Figure 6 plots the mean word recognition score of each of the two tests across conditions.

Table 5

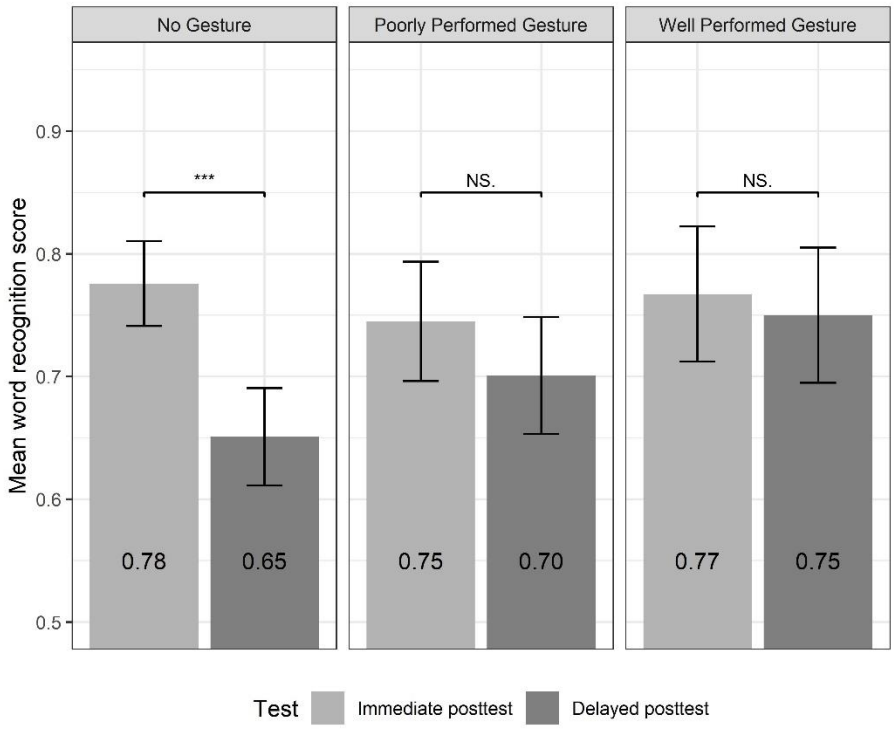
Means and Standard Deviations of Word Recognition Score Across Test and Group

Condition	Immediate posttest	Delayed posttest
	M (SD)	*M (SD)*
NG	0.78 (0.33)	0.65 (0.38)
PPG	0.75 (0.35)	0.70 (0.35)
WPG	0.77 (0.33)	0.75 (0.33)

Note. NG = No Gesture; PPG = Poorly Performed Gesture; WPG = Well performed Gesture.

Figure 6

Mean Word Recognition Score



Note. The numbers labelled on the bars represent the mean score. Error bars indicate 95% confidence interval. $^{***}p < .001$; $^{**}p < .01$; $^{*}p < .05$; NS. = not significant.

3.4 Discussion and Conclusion

The present study examined whether producing visuospatial hand gestures depicting L2 consonantal aspiration would help Catalan speakers learn the pronunciation of these non-native sounds, as well as the meaning of words containing such contrast. With respect to the first research question, the results revealed that even though producing visuospatial gestures cueing the air burst of aspirated plosives had limited effects on L2 speech perception, it clearly improved the production of aspirated plosives right after training. Importantly, when the learners appropriately performed the gestures during training, such positive effects could be observed not only immediately after training but also three days later. As for the second research question, the results showed that gestural training significantly retained the accuracy of word recognition as opposed to non-gestural training, regardless of how accurately the learners performed the gestures. Taken together, our results confirmed the benefits of gesture production in L2 pronunciation learning and pointed out the importance of gesture performance.

3.4.1 Effects of visuospatial hand gestures on the perception of L2 segmental features

The results of the identification task showed that participants in all four conditions improved their identification score after a short training session yet did not manage to maintain this gain over three days. Thus, performing gestures, be they well performed or not, did not reveal more benefits than non-gestural training. This seems to be in line with several previous studies showing that hand gestures have limited effects on the perception of novel phonological contrasts related to duration (Hirata et al., 2014; Hirata & Kelly, 2010; Kelly et al., 2017; P. Li, Baills, et al.,

2020) or aspiration (Xi et al., 2020). However, other studies have detected positive effects of hand gestures on the perception of L2 lexical tones (Baills et al., 2019; Morett & Chang, 2015; Zhen et al., 2019). It might well be that the identification task for assessing the perception of the two-way contrast was relatively easy and yielded high mean scores already at pretest; little room was left for improvement. Hence, future studies should employ more demanding tasks that will therefore be more discriminating.

3.4.2 Effects of visuospatial hand gestures on the production of L2 segmental features

The acoustic analyses of the VOT of the aspirated plosives in the imitation task showed clear evidence that appropriate embodied training favors the production of the non-native consonants. First, while the No Gesture and Well Performed Gesture groups produced more accurate VOT at the immediate posttest, the No Gesture group showed a significant decrease at delayed posttest, which is not the case of the Well Performed Gesture group. By contrast, the Poorly Performed Gesture group did not show any change in VOT ratio. In other words, inappropriately performing hand gestures led to null effects on VOT, and not performing gestures could not maintain the improvement. As for the maintenance of training effects, although a medium effect size from pretest to immediate posttest was obtained in both groups (No Gesture: d = 0.49, and Well Performed Gesture: d = 0.69), when comparing the delayed posttest to the pretest, the No Gesture group showed a small effect size (d = 0.22), whereas the Well Performed Gesture group still maintained the same effect size (d = 0.69). This suggests that the Well Performed Gesture group

was more stable than the No Gesture group in maintaining the training effects.

Regarding the overall pronunciation, the group of participants who appropriately performed hand gestures during training obtained the best learning outcome. First, the fact that all experimental groups improved immediately after training, whereas no significant gain was observed in the Control group, points to the conclusion that the different training methods accounted for the significant changes at the two posttests. Interestingly, even though the Poorly Performed Gesture group showed a significant improvement at the immediate posttest, it was with a small effect size (d = 0.36), suggesting that not being able to perform hand gestures appropriately triggers limited effects. More importantly, the difference between the three groups can be observed in their performance in the delayed posttest. While the Poorly Performed Gesture group did not maintain the training effects after three days, the No Gesture group did so. However, the Well Performed Gesture group showed continuous improvement in their mean pronunciation score over the three tests. More importantly, while the Well Performed Gesture group approached a medium-level effect size (d = 0.59), the No Gesture group only displayed a small-level effect size (d = 0.22), which was smaller than that from pretest to immediate posttest (d = 0.33).

Both the perceptual ratings and the acoustic analyses illustrate the sharp asymmetry between outcomes for the Well Performed Gesture and the Poorly Performed Gesture groups, pointing to the importance of assessing gesture performance during embodied training, something which has been largely neglected in previous research. The poor performance

of hand gestures might be a signal for cognitive overload. Adding multimodal information (e.g., verbal instruction, gesture production, etc.) might increase the cognitive load of participants who cannot handle the task due to low language skills (Kelly & Lee, 2012) or task difficulties (Post et al., 2013). Thus, embodied training methodologies should take into account gesture accuracy and task effects, making sure that participants can handle the tasks.

The results on the identification task contrasted with those of the imitation task. This is in line with previous studies showing that when novel contrasting features are being learned, an improvement in speech perception does not necessarily go together with an improvement in speech production (Nagle, 2018) and that the two modalities may be somewhat independent of each other (Baese-Berk & Samuel, 2016), especially when learners are at an early stage of L2 acquisition (Zampini, 1998). Moreover, according to the transfer-appropriate processing principle, knowledge is more easily recalled when the retrieval shares similar cognitive processes with the training (e.g., Franks et al., 2000; Lightbown, 2008; Morris et al., 1977; Segalowitz, 1997, 2000). In other words, the fact that the pronunciation training was provided through an oral imitation task might explain why the training effects were more effective in the imitation task than in the identification task. In addition, when people acquire skills through practice, the effect of practice is skill-specific (DeKeyser, 2015). Since our training focused on production patterns and not on perception, this might have led to specific gains in pronunciation skills.

Finally, the lack of significant main effect of familiarity (trained vs. untrained items) or its interaction with aspiration, test, and gesture performance indicates that the generalization occurred from trained items to untrained items regardless of the training method. In other words, participants' responses in the perception and production were not affected by whether or not a particular item was presented in the training session.

3.4.3 Effects of visuospatial hand gestures on L2 vocabulary learning

The results of the word-meaning association task revealed positive effects of gestural training on maintaining the accuracy of word recognition. While the three groups obtained similar scores at the immediate posttest, there was a general decay after three days. However, significant decay was only observed in the No Gesture group. These results complement recent studies reporting positive effects of gestures encoding suprasegmental features on L2 vocabulary learning (Baills et al., 2019; Kushch et al., 2018; Morett & Chang, 2015) by showing a similar effect related to segmental features. That is, even though gestures cue important phonetic features, they can strengthen the link between phonological forms and semantic meaning.

Interestingly, it seems that learners' gesture performance was not as relevant for learning vocabulary meanings as it was for learning pronunciation. This might be due to the fact that the target words in the vocabulary training were phonologically easier to process than those in the pronunciation training (i.e., disyllabic vs. monosyllabic words). However, embodied training with gesture production did help learners to maintain their recognition of newly learned words bearing aspiration contrasts. Given that iconic hand gestures conveying semantic information

might impair the learning of minimal word pairs (i.e., words with high phonological demands) (Kelly & Lee, 2012), visuospatial hand gestures conveying phonological information of difficult phonemes may help better process the same type of minimal pairs. These visuospatial gestures may offload the phonological demands to the visual channel and save participants' cognitive sources, which can be allocated to phonological and lexical recall.

3.4.4 Limitation

The current study has several limitations. First, although we hypothesized that poor gesture performance could have been due to cognitive overload, we did not measure cognitive load. Future investigations could include complementary measures in this respect (see Brünken et al., 2003; Skulmowski & Rey, 2017 for measurement of cognitive load). Second, it might well be that learners' gesture performance accuracy was related to their efforts to the learning, the comfort they felt when making gestures, etc. These individual factors were not assessed in the current study but should be explored in more depth in future studies.

To conclude, the present study shows that actively producing hand gestures may play a positive role in producing novel consonants and the learning of novel words bearing these consonants. Importantly, our results show that the accuracy with which learners perform those gestures during training is an essential issue for multimodal training and that teachers need to pay attention to this issue and design tasks that are adequate for the learners' needs. All in all, this study expands our understanding of embodied cognition by providing direct evidence for the positive role of gestures in the field of L2 pronunciation training.

4

CHAPTER 4: EMBODIED PROSODIC TRAINING HELPS IMPROVE NOT ONLY ACCENTEDNESS BUT ALSO VOWEL ACCURACY

Li, P., Xi, X., Baills, F., Baqué, L., & Prieto, P. (under review). Embodied prosodic training helps improve not only accentedness but also vowel accuracy. *Language Teaching Research.*

4.1 Introduction

In recent years, the importance of pronunciation instruction in L2 teaching and learning has received increasing attention (e.g., Kang et al., 2019; Lee et al., 2015; Saito & Plonsky, 2019). Recent studies suggest that both suprasegmental (or prosodic) and segmental features should be trained in pronunciation instruction (e.g., J. Lee et al., 2015) and that teachers should take advantage of the strong relationship between the two (X. Wang, 2020; Zielinski, 2015). This is in accordance with results suggesting that even though suprasegmental features (e.g., Anderson-Hsieh et al., 1992; Trofimovich & Baker, 2006) and segmental accuracy (e.g., Isaacs & Trofimovich, 2012; Saito et al., 2016, 2017) are both important for native judgment of accentedness. In addition, suprasegmental features and fluency measures influence comprehensibility (e.g., Crowther et al., 2016; Isaacs & Trofimovich, 2012; Saito et al., 2017), while only segments with high functional load seem to affect comprehensibility (Munro & Derwing, 2006; Suzukida & Saito, 2019).

Notwithstanding, most research on L2 pronunciation has focused on specific phonemes in order to improve learners' pronunciation proficiency (e.g., Hoetjes & van Maastricht, 2020; Lambacher et al., 2005; Lord, 2005, 2008; K. Saito & Munro, 2014; Xi et al., 2020), little is known about the potential effects of prosody-based pronunciation training on the production of non-native segmental features. The main goal of the

present study will be to empirically assess whether embodied prosodic training (i.e., the use of visuospatial hand gestures by instructors to highlight the melodic and rhythmic features of sentences) can have beneficial effects at the segmental level.

4.1.1 Prosodic training and L2 pronunciation instruction

Previous research has looked at the effects of prosodic pronunciation instruction (i.e., focusing on suprasegmental features like speech rate, rhythm, intonation, etc.) compared to segmental pronunciation instruction (i.e., focusing on specific vowels and consonants) in the L2 classroom on learners' pronunciation proficiency. It has been shown that prosodic training may produce stronger gains than segmental training (Derwing et al., 1998; Gordon et al., 2013; Gordon & Darcy, 2016; Y. Saito & Saito, 2017; R. Zhang & Yuan, 2020). Derwing et al. (1998) compared the two types of instructions and found that though both training methods were effective in controlled and spontaneous speech production, only prosodic training improved comprehensibility and fluency in free speech. Likewise, Gordon et al. (2013) and Gordon and Darcy (2016) found that prosodic training enhanced learners' comprehensibility, whereas segmental training did not. Later, Y. Saito and Saito (2017) confirmed that prosodic training could improve comprehensibility and the production of non-native intonational patterns. Furthermore, a recent

study (R. Zhang & Yuan, 2020) found that suprasegmental training triggered delayed positive effects on comprehensibility in spontaneous speech compared to segmental training and non-specific pronunciation training.

However, most of the studies generally assessed overall pronunciation or pronunciation at suprasegmental level (except Gordon & Darcy, 2016; Y. Saito & Saito, 2017), little is known about whether prosodic training may also help improve the pronunciation of segmental features.

To our knowledge, only a handful of empirical studies have explored the potential effects of L2 prosodic training on the improvement of non-native segmental features, with mixed results (e.g., Gordon & Darcy, 2016; Hardison, 2004; Missaglia, 2007; Y. Saito & Saito, 2017). Focusing first on the positive results, Missaglia (2007) showed that training in prosody could yield greater gains in overall pronunciation and segmental accuracy than training in segments. Likewise, Hardison (2004) confirmed that the gains in prosodic accuracy obtained from intonation training could be generalized to segmental accuracy as well. More recently, Y. Saito and Saito (2017) found that suprasegmental training involving intonation, rhythm, and speech rate helped improve Japanese students' vowel accuracy in L2 English production. According to the authors, this effect seemed to stem from the students' gains in accurately reproducing rhythmic structures. By contrast, some research failed to find beneficial

effects of prosodic training on L2 segmental accuracy. Gordon and Darcy (2016) reported that, although suprasegmental training helped improve speech comprehensibility, it led to no improvement in the pronunciation of L2 vowels.

There are several reasons to hypothesize that prosody may help improve the pronunciation of segments. First, prosodic and segmental structures are two integrated and interdependent components in producing a language. Following up on the Prosodic Bootstrapping hypothesis, which postulates that prosodic features (e.g., rhythm, tempo, pitch) may help bootstrap syntactic and lexical features in early first language acquisition (Christophe et al., 1997, 2008), and given the strong interdependence between prosodic and segmental structure, we hypothesize that prosody can bootstrap the pronunciation of segments. Importantly, recent evidence has provided some results showing the bootstrapping effect of rhythmic training in improving the imitation abilities in an L2 (Campfield & Murphy, 2014). Interestingly, this idea has been applied to speech therapy (Bedore & Leonard, 1995). Moreover, regarding the interaction between prosody and segments, the Sonority Expansion hypothesis (Beckman et al., 1992) holds that speech prominence can make the vowel more opened, while the Hyperarticulation hypothesis (de Jong, 1995) claims that despite openness, speech prominence may even affect the lip roundedness and blackness of the vowels. In addition, a non-

prominent position, in general, may compress the pronunciation of segments in duration and formant frequencies (Walker, 2011, p. 16).

The mechanisms underlying the effects of prosodic training on segmental accuracy seem to be linked to how such training enhances sensitivity to rhythmic patterns, as noted above with reference to Y. Saito and Saito (2017). Similarly, during the suprasegmental training described by Missaglia (2007), because Italian learners of German were asked to exaggeratedly produce only one stressed syllable in each sentence, all remaining syllables were reduced so that the target vowel reduction would naturally occur. In other words, these studies took advantages of the interaction between rhythmic structure and vowel quality. Moreover, in highlighting the suprasegmental features, many studies have found that hand gestures mimicking the target suprasegmental features may be of help. In the next section, we motivate the use of gestural cues to highlight suprasegmental features, namely, the embodied approach to pronunciation training.

4.1.2 Embodied approaches to training L2 pronunciation

Embodied Cognition (EC) captures the strong relation between mind and body (Ionescu & Vasc, 2014), which holds that the body is tightly involved in human cognitive processes (Barsalou, 2008; Foglia & Wilson, 2013) and may therefore have a strong impact on learning and education (e.g., Kiefer & Trumpp, 2012; Shapiro & Stolz, 2019). The cognitive

offloading theory holds that people tend to offload their cognitive load onto the environment or the body to reduce the occupation of working memory or attention abilities (Risko & Gilbert, 2016). These theories thus have important implications for education, in that "embodiment offers either a causal route to more effective learning or a diagnostic tool for measuring conceptual understanding" (Shapiro & Stolz, 2019, p.30).

Like many areas of learning, where the Embodied Cognition paradigm has been extensively applied, second language teachers make frequent use of embodied strategies in their classrooms. They use hand gestures, tactile information, hand-clapping, and tapping to illustrate various aspects of pronunciation, including syllabification, word stress, rhythm (Smotrova, 2017), difficult phonemes (Rosborough, 2010), and segmental features (Hudson, 2011).

Notably, a growing body of empirical research has shed light on the positive role of embodied training on L2 pronunciation in both suprasegmental and segmental domains. It has been documented that beat gestures highlighting speech prominence may benefit L2 pronunciation (Gluhareva & Prieto, 2017; Kushch, 2018; Llanes-Coromina et al., 2018). Similarly, clapping hands or tapping fingers to the rhythm of words has been shown to benefit L2 pronunciation (Baills & Prieto, 2021; B. Lee et al., 2020; Y. Zhang et al., 2020). Moreover, hand gestures tracing pitch contours in space performed over the nuclear-accented syllable were

found to favor the pronunciation of L2 intonational patterns (Yuan et al., 2019). Likewise, illustrating vowel length by short and long horizontal hand sweep gestures can also improve the pronunciation of the non-native short and long vowels (P. Li, Baills, et al., 2020). Finally, embodying phonetic features can also help the pronunciation of non-native segmental features, like aspiration (Amand & Touhami, 2016; P. Li et al., 2021; Xi et al., 2020), interdental consonants (Hoetjes & van Maastricht, 2020; Ozakin et al., under review) and vowels (Hoetjes & van Maastricht, 2020).

In practice, various types of teaching techniques involving embodiment have been proposed, among them the so-called verbotonal method (e.g., Guberina, 2008; Intravaia, 2000; Renard, 1979) has drawn much attention. It encourages the combination of prosody and body movements like hand gestures for phonetic corrections at both the segmental and supra-segmental levels. For example, to trigger a more target-like pronunciation of the French front rounded /y/ for Spanish speakers (who often pronounce it as /u/), teachers may place the /y/ in a rising intonation contour (Renard, 2002), embody the rise with an upward hand gesture (Billières, 2002), and put it in various prosodic positions in meaningful discourses (Wlomainck, 2002).

Nevertheless, only a few studies have empirically assessed the role of the verbotonal method in actual teaching practice, and they have yielded inconclusive findings. Whereas Alazard et al. (2010) found that the verbotonal method could improve L2 French learners' fluency in oral reading, and Alazard (2013) reported that this method might particularly benefit beginning learners' speech fluency, in a more recent study, Alazard-Guiu et al. (2018) reported that the verbotonal method could not outperform the traditional focus-on-form training in improving segmental accuracy. Nonetheless, given the limited sample size in this last study (eight participants), one might expect more conclusive results in experiments with larger populations.

4.1.3 The present study

To assess the value of embodied prosodic training strategies on L2 pronunciation, and especially its effects at the segmental level, the present study will investigate whether intermediate Catalan learners of French can benefit from embodied prosodic training to boost their reading pronunciation, with a focus on non-native front rounded vowels.

Catalan learners of French face clear challenges in the acquisition of non-native segmental and suprasegmental patterns. At the segmental level, the French front rounded vowels /y, ø, œ/ contrast with the back rounded vowels /u, o, ɔ/ (Darcy et al., 2012), whereas only back rounded vowels /u, o, ɔ/ are part of the Catalan vocalic system (Wheeler, 2005). Based

on the observation that learners of French whose native languages do not have front rounded vowels in their vocalic inventories tend to assimilate the three vowels to their back counterparts (see Darcy et al., 2012; Levy & Law, 2010 for English speakers; Racine & Detey, 2019 for Spanish speakers; Hannahs, 2007 for a review), we hypothesize that Catalan speakers may also display such assimilation, whereby the front rounded /y, ø, œ/ are produced as their back rounded counterparts /u, o, ɔ/ respectively. Regarding suprasegmental features, French stress is assigned at the phrase level (i.e., Accentual Phrase, or AP), marked by a phrase-initial optional high tone and a phrase-final obligatory high tone, implying that stress is a demarcative property of the AP rather than the word (e.g., Fougeron & Jun, 2002; Jun & Fougeron, 2000). Contrastingly, Catalan does not show evidence for AP, whereas the intermediate phrase generally consists of more than one prosodic word (Prieto et al., 2015). At the level of Intonational Phrase (IP), although both languages have a final nuclear prominent accent in the last content word of the IP (Delais-Roussarie et al., 2015; Prieto et al., 2015), realized by longer duration compared to unstressed syllables, the durational ratio is larger in French than in Catalan (Baills & Prieto, 2021). Therefore, these differences in prosody may have influences on the overall pronunciation proficiency.

Hence, two main research questions will be addressed, as follows:

RQ1: Does the embodied prosodic training improve Catalan learners' overall pronunciation proficiency of French more than comparable non-embodied training? To answer this question, we will compare the effects of non-embodied and embodied training on accentedness, comprehensibility, and fluency.

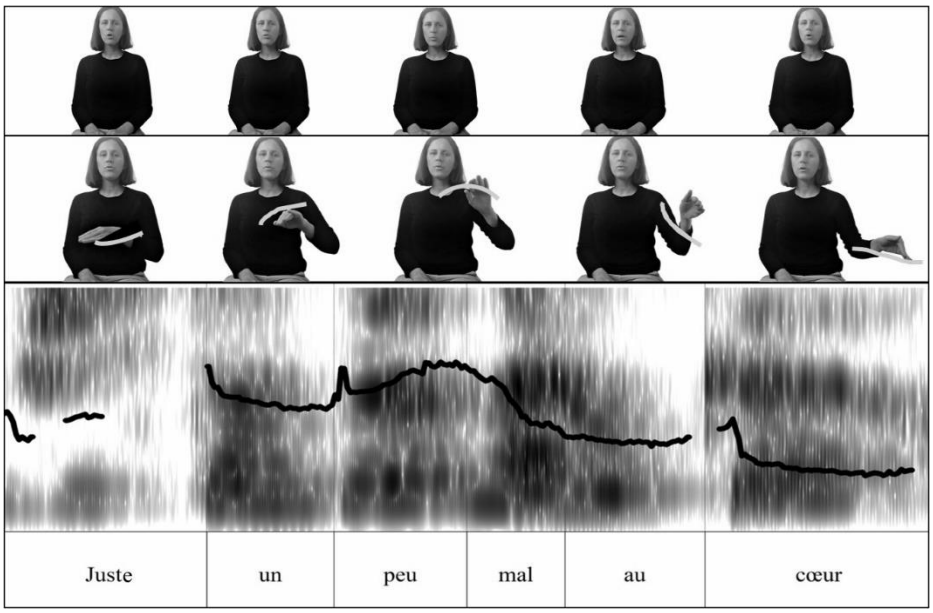
RQ2: Does exposure to embodied prosodic training benefit Catalan learners' pronunciation of non-native vocalic features more than comparable non-embodied training? As mentioned above, the target phonemes will be the non-native front rounded vowels /y, ø, œ/ of French.

To implement the embodied prosodic training, we used hand gestures to mimic pitch and durational features. Figure 1 illustrates one of the instructor's hand movements illustrating the prosodic structure of the sentence *J**u**ste un p**eu** mal au c**œu**r*. 'Just a bit sick at heart.' The up and down hand movements not only show pitch peaks and valleys but also indicate speech prominence, thus highlighting rhythmic and intonational structures. The rightmost images show how phrase-final lengthening patterns are also embodied through the prolongation of the instructor's horizontal hand movement. Note that the target front rounded vowels (boldface in the sample sentence above) appear in different positions in the utterance.

In addition, we assessed participants' pronunciation proficiency via two complementary tasks, namely sentence imitation task and discourse reading task. In particular, imitation tasks do not seem to necessarily reflect the productive knowledge of the difficult phonemes in an L2 (Llompart & Reinisch, 2019). Therefore, in order to draw a more complete picture of the French pronunciation of the participants, we decided to involve the two tasks, which allows auditory input (imitation) and orthographic input (reading) on different levels (sentence and dialogue).

Figure 1

A sequence of images illustrating the visuospatial hand gestures performed by the instructor as she produces a sample sentence in the non-embodied (upper panel) or embodied condition (lower panel)



4.2 Methods

4.2.1 Participants

Fifty-seven undergraduate students pursuing degrees in translation or applied languages (53 females, aged 18–46 years, M_{age} = 19.89 years, *SD* = 3.63) were recruited in spring 2020 from a public university. The participants considered themselves Catalan-Spanish bilinguals and reported using Catalan for their daily verbal communication on average 62.81% of the time (*SD* = 29.67). Prior to enrolling in this study, all participants signed a consent form which gave the researchers permission to collect and analyze the audio and video recordings obtained during the experiment.

The pronunciation training session was incorporated into two French language courses, intended for first- and second-year students, respectively. Participation was mandatory. We took advantage of the fact that French courses were divided into two groups for special speaking practice sessions each week, and thus for each course, one practice group received embodied prosodic training (*n* = 28; first-year = 9; second-year = 19) while the other group received non-embodied training (*n* = 29; first-year = 7; second-year = 22).

4.2.2 Materials

The experiment was a between-subject training study with a pretest/post-test/delayed posttest paradigm. In this section, we describe the materials created for the experiment, including the audio-visual stimuli for the training sessions as well as the auditory and textual stimuli for the three tests.

a) Audio-visual stimuli for the pronunciation training sessions

Training dialogues. The training materials were adapted from a French pronunciation textbook which provided a series of dialogues featuring two interlocutors experiencing interesting situations (Martinie & Wachs, 2006). For this experiment, we selected three dialogues designed to train the three front rounded vowels, namely, /y, ø, œ/, and modified the content to increase their overall frequency (see Table C1).

Table 1 summarizes the prosodic positions of the three target vowels in the three training dialogues. More than half (57%) of the target vowels were pitch-accented, and of those, most (59%) carried a high tone or were in a rising intonation. It should be noted that since many functional words contain /y/ (e.g., *tu* 'you', *du* 'of the', *une* 'a', etc.), the frequency of /y/ is inevitably higher than that of /ø/ and /œ/ and functional words are often unaccented.

Table 1

Count and Proportion of the Target Front Rounded Vowels separated by Prosodic Patterns in the Training Dialogues

	/y/	/ø/	/œ/
Prosodic Pattern	*n* (%)	*n* (%)	*n* (%)
Unaccented	28 (53%)	10 (50%)	0 (0%)
Accented	25 (47%)	10 (50%)	16 (100%)
High/rising	14 (56%)	7 (70%)	9 (56%)
Low/falling	11 (44%)	3 (30%)	7 (44%)

Audio-visual training stimuli. The audio-visual stimuli consisted of two parts: three enactments of the dialogues and the sentence-by-sentence training clips with or without gestures.

The enactments of the three dialogues were performed in pairs by four female amateur actors (all native speakers of French). Each performance lasted around 45 seconds.

For the sentence-by-sentence training clips, two experienced female teachers of French with native French proficiency were video-recorded, producing each of the sentences of the three training dialogues. Before recording, the two instructors watched the three enactments of the dialogues as many times as necessary and imitated the speech of the actors. They were filmed against a white wall, with their face and the upper half of their body visible. For the non-embodied condition, the instructors

produced the sentences naturally without any body movements other than those strictly related to oral articulation. For the embodied condition, they produced each sentence along with the visuospatial hand gestures to visually illustrate the prosodic information of the sentence (see Figure 1 for an example) while keeping the rest of the body still. They were additionally provided with pictures generated with Praat (Boersma & Weenink, 2020) displaying the pitch contour as a curved line and the segmentation of each word in the sentence so that they could trace the pitch track by hand movements. The appropriateness of the visuospatial gestures was checked by comparing them to the visual intonation patterns of the dialogues generated by Praat. Additionally, in order to avoid any potential differences in speech across the two conditions, the audio tracks of the gesture videos were added to the corresponding no-gesture videos, replacing the original audio recording. All sentences were recorded four times, and the authors selected the best version of each sentence so that a total of 128 video clips were selected, 64 for each condition. With these materials, six training videos were created (3 dialogues × 2 conditions). Importantly, for each dialogue, each instructor was assigned a separate role so that the two instructors spoke in alternating turns.

b) Pretest and posttest materials

For the dialogue-reading task, in addition to the three dialogues that had been trained, a fourth untrained dialogue was selected from the same textbook (Martinie & Wachs, 2006). Like the trained dialogues, the untrained dialogue was adapted to increase the frequency of the target front rounded vowels (see Table C1, untrained dialogue).

The stimuli for the sentence imitation task were 15 sentences selected from the dialogue-reading task, 12 sentences from the trained dialogues, and three sentences from the untrained dialogue (see Table C2). The sentences were audio-recorded by the four actors who performed the dialogues. Each sentence was read by the person who said it in the video. The three untrained sentences were read by the second author.

c) French language proficiency and prior learning experience

Since the participants were not necessarily beginning learners of French when they were admitted to the undergraduate program, we asked them to answer a questionnaire regarding their French language learning background in terms of age of onset learning, years of formal learning, months of study abroad in a French-speaking country, and months of extracurricular courses. In addition, they were asked to self-assess their

French proficiency from 1 (A1) to 6 (C2) according to the CEFRL (Common European Framework of Reference for Languages, see Council of Europe, 2001).

4.2.3 Procedure

Due to the impact of the COVID-19 pandemic, all spring 2020 courses were conducted on-line, so testing and training for this study also took place on-line. While the training sessions were carried out online under the supervision of the teachers during the periods designated for speaking practice, the testing sessions were carried out individually as course homework.

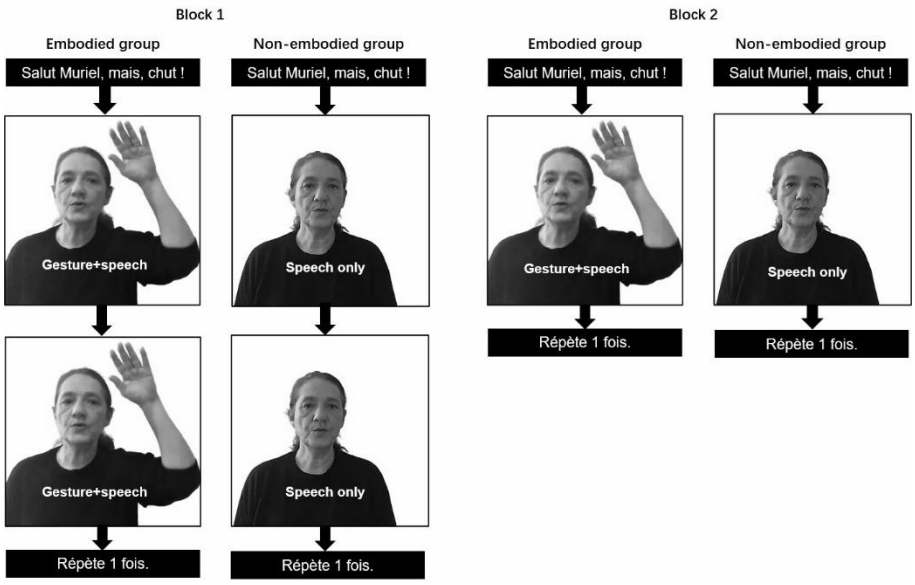
The experiment lasted six weeks. In the first week, the participants performed the pretest tasks, which consisted of the dialogue-reading and sentence imitation tasks. In the dialogue-reading task, participants had to read aloud the text of the dialogues presented online using Alchemer software (https://www.alchemer.com/). In the sentence imitation task, participants listened to each of the 15 French sentences once and imitated each of them immediately afterwards. The presentation of the sentences was randomized automatically. In both tasks the participant's voice was automatically recorded by an on-line camera (http://webcamera.io).

From the second week to the fourth week, participants received three sessions of audio-visual training, one session per week. Each training

session followed exactly the same procedure. First, the seminar teacher explained the words that might be unfamiliar to the participants in the training materials. Then participants watched the video and performed the training. Specifically, they first watched the enactment of the dialogue to be trained (45 s) and then read the dialogue script aloud by themselves (90 s). Following this, the dialogue was trained sentence by sentence, in the order they appeared in the dialogue. Each sentence was trained in two blocks. In the first block participants first saw the sentence written in French (5s); then, one of the two instructors read the sentence twice with or without gestures, depending on the condition. After that, a black screen displaying "Repeat that once" appeared (5 s), allowing the participants to repeat the sentence once. In the second block, however, the instructor uttered each sentence only once, while the rest of the procedure remained the same as that of the first block (see Figure 2).

Figure 2

Audiovisual Training Procedure in Two Blocks, and Across Conditions

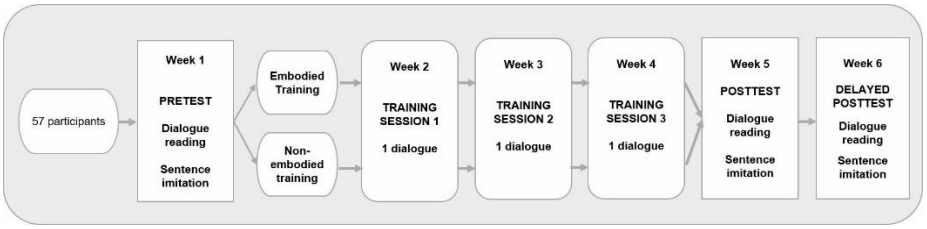


After the audio-visual training, participants were asked to watch the dialogue enactment a second time (45 s) and then read aloud the script of the dialogue again (90 s). In total, each training session lasted around 15 minutes.

In the fifth and sixth weeks, the participants took the posttest and delayed posttest, respectively, in which they repeated the dialogue-reading and sentence imitation tasks as in the pretest. The procedure used for the two tasks was exactly the same as in the pretest. It is important to note that the interval between two training sessions and/or tests was set at one week. The experimental procedure is schematized in Figure 3.

Figure 3

Experimental procedure



4.2.4 Data Coding

The participants' pronunciation of the target dialogues and sentences produced in the two tasks were assessed both perceptually and acoustically by means of formant analysis of the target front rounded vowels. A total of 684 recordings (4 dialogues × 3 tests × 57 participants) were obtained from the dialogue-reading and 2,565 recordings (15 sentences × 3 tests × 57 participants) from the sentence imitation task.

a) Pronunciation assessment

Three native French-speaking teachers (1 female, aged 29–39 years, M_{age} = 35 years) rated all the recordings produced by the participants. The three raters had taught French in a Catalan-speaking city for at least three years by the time of recruitment. On a five-point Likert scale (1 = not at all; 5 = completely), raters' self-evaluation of familiarity with Catalan-accented French yielded an average of 4.6 and self-reported knowledge in French phonology an average of 4, which suggests that they were very sensitive to Catalan-accented French and had enough linguistic knowledge to give reliable judgments.

Before performing the rating task, all raters were trained in a 45-minute session to familiarize them with the evaluation system. For the dialogue-reading task, they rated accentedness, comprehensibility, and fluency, while for the sentence imitation task, they only rated accentedness, given that the sentences were very short (around 2 s). Raters were asked to listen to each recording and rate each measure on a scale from 1 to 9. Following Munro and Derwing (2015), accentedness measured "the difference in pronunciation as compared with the native speakers" (1 = "very strong foreign accent", 9 = "no foreign accent at all"); comprehensibility was defined as "the degree of difficulty in understanding the speech" (1 = "incomprehensible", 9 = "completely comprehensible") while fluency referred to the "fluidity of speech" (1 = "disfluent", 9 = "fluent"). The ratings were performed online in 21 one-hour batches, each of which contained 57 dialogues or 285 sentences. The rating was paced at one batch per day over 21 days. All the recordings were presented to the raters randomly through the online software Alchemer.

Inter-rater reliability was checked by a series of two-way mixed, consistency, average-measures (k = 3) Intra-Class Correlation (ICC) analyses (Hallgren, 2012). The ICC and their 95% confidence interval were calculated using the *irr* package version 0.84.1 (Gamer et al., 2019) in R version 4.0.2 (R Core Team, 2014). The three raters showed a good level of agreement in the rating of accentedness (ICC = 0.81, 95% CI [0.79, 0.84]), comprehensibility (ICC = 0.83, 95% CI [0.80, 0.84]), and fluency

(ICC = 0.78, 95% CI [0.75, 0.81]) for the dialogue-reading task, as well as a good level of agreement in the accentedness rating (ICC = 0.81, 95% CI [0.79, 0.82]) for the sentence imitation task (see Koo & Li, 2016 for the interpretation of ICC scores). Therefore, the ratings of the three raters of each measure were averaged per each item for each participant in order to create the scores for accentedness, comprehensibility, and fluency to be analyzed.

b) Acoustic analyses

In the field of acoustic phonetics, formant analysis is often employed to capture vowel quality differences, and the first two formant frequencies (i.e., F1 and F2) are often used to describe the vowels in terms of tongue height (mouth aperture) and tongue frontness/backness (tongue position) (Johnson, 2011). In order to minimize gender influence on formant frequencies, for the current analysis we only included female participants (N = 53). This did not substantially reduce our sample size, given that there were only two male participants in each condition. In order to ensure that the acoustic analyses were comparable between the two tasks, the first author annotated exactly the same front rounded vowels that the female participants produced in both the dialogue-reading task and the sentence imitation task using Praat software. Accordingly, for each task, a total of 3,021 tokens (19 tokens × 53 participants × 3 tests) were annotated. After annotation, the mean F1 and F2 values of each token were

extracted from Praat. The acoustic data were then transformed from Hertz to Bark to normalize the individual differences in vocal tract length using the following formula: $Bark = 7ln\{(Hz/650) + [(Hz/650)^2 + 1]^{½}\}$ (Traunmüller, 1990, see Gordon & Darcy, 2016; K. Saito & Munro, 2014 for similar decisions).

4.2.5 Statistical analyses

Sixteen Generalized Linear Mixed Models were applied to the following outcome measures using the *glmmTMB* package, version 1.0.2.1 (Brooks et al., 2017) in R, version 4.0.2. For the dialogue-reading task, three averaged perceptual measures (accentedness, comprehensibility, and fluency) and six acoustic measures (the Bark normalized F1 and F2 values of /y, ø, œ/) were used, while for the sentence imitation task, the accentedness score and six acoustic measures were used. For all models, the fixed factors were condition (two levels: non-embodied and embodied), test (three levels: pretest, posttest, and delayed posttest), and their interaction. Another possible fixed factor was training (trained items vs. untrained items). However, an initial analysis revealed that none of the sixteen models obtained the expected three-way interaction of Training × Test × Condition. This, together with the fact that only three models obtained a significant main effect of training, and the lack of balance between trained and untrained items, led us to exclude this factor, so the analysis was nested to our research purpose. Two random intercepts were

included, one for participants, the other one for items. Significance was determined by Type II Wald *chi*-squared tests using the *car* package, version 3.0.9 (Fox & Weisberg, 2019). The post-hoc analyses were a series of Bonferroni pairwise comparisons performed with the *emmeans* package, version 1.4.8 (Lenth et al., 2020). Effect size was assessed using Cohen's *d* in the post-hoc analyses (small: $d \geq 0.2$; medium: $d \geq 0.5$; large: $d \geq 0.8$, see Cohen, 1988).

4.3 Results

4.3.1 Homogeneity across groups

A series of Whitney-Mann *U*-tests were run to check if there were any differences in French proficiency or prior learning experiences across groups. Although participants displayed large individual differences within groups, no statistical differences were found between groups in terms of age, age of onset learning, years of formal learning, months of study abroad, months of extracurricular courses, or self-assessed proficiency (all $p > .05$, see Table 2 for descriptive and inferential statistics). These results suggest that the two training groups were comparable in terms of French learning experience and proficiency level.

Table 2

Means, Standard Deviations, and Mann-Whitney U Test Results of Individual Differences in French Learning Experience and Self-Assessed Proficiency Across Groups

	Non-embodied		Embodied		*U*	*p*
	M	*SD*	*M*	*SD*		
Age	20.41	4.98	19.36	1.03	327.00	.174
Age of onset learning	13.72	3.06	14.11	2.99	413.00	.910
Formal learning (years)	5.52	3.01	5.18	2.57	381.00	.687
Study abroad (months)	0.45	1.24	0.71	1.43	473.50	.142
Extracurricular courses (months)	1.97	3.09	1.61	2.85	385.00	.721
Self-estimated proficiency [a]	3.07	1.00	3.25	0.93	434.00	.637

[a] Self-estimated proficiency was evaluated from 1 (A1) to 6 (C2). Thus an average score of around 3 indicates that the group of participants showed an overall B1/intermediate level.

4.3.2 Dialogue-reading task

a) Accentedness comprehensibility and fluency

The descriptive data for the three perceptual measures are displayed in Table 3.

Table 3

Means and Standard Deviations of the Accentedness, Comprehensibility, and Fluency Scores Across Group and Test for the Dialogue-Reading Task.

	Non-embodied		Embodied	
	M	*SD*	*M*	*SD*
Accentedness				
Pretest	4.41	1.39	4.53	1.35
Posttest	4.64	1.32	4.85	1.17
Delayed posttest	4.70	1.34	5.07	1.18
Comprehensibility				

Pretest	5.33	1.43	5.44	1.45
Posttest	5.80	1.41	5.82	1.21
Delayed posttest	5.92	1.32	5.92	1.29
Fluency				
Pretest	5.18	1.19	5.32	1.61
Posttest	5.94	1.31	6.09	1.28
Delayed posttest	6.08	1.23	6.27	1.21

In what follows, we report the results of the statistical analyses for each dependent variable separately.

Accentedness. Analysis of accentedness scores revealed a significant main effect of test ($\chi^2(2) = 68.05$, $p < .001$) as well as a significant two-way interaction of Test × Condition ($\chi^2(2) = 6.41$, $p = .041$), although no significant main effect of condition ($\chi^2(1) = 0.56$, $p = .453$) was found. The two-way interaction indicates that the accentedness score varied across condition and test. The post-hoc results are as follows. For the non-embodied group, there was a significant improvement from pretest to posttest ($\Delta = 0.23$, $SE = 0.07$, $t(675) = 3.27$, $p = .003$, $d = 0.43$) and this improvement was maintained at delayed posttest, indicated by a significant improvement from pretest to delayed posttest ($\Delta = 0.29$, $SE = 0.07$, $t(675) = 4.04$, $p < .001$, $d = 0.53$). By contrast, the accentedness score progressively improved in the embodied group across the three tests, indicated by a significant improvement from pretest to posttest ($\Delta = 0.32$, $SE = 0.07$, $t(675) = 4.38$, $p < .001$, $d = 0.59$), and from posttest

to delayed posttest (Δ = 0.22, *SE* = 0.07, *t*(675) = 3.07, *p* = .007, *d* = 0.41).

Comprehensibility. Analysis of comprehensibility scores revealed a significant main effect of test ($\chi^2(2) = 101.58$, $p < .001$), but no significant effects of condition ($\chi^2(1) = 0.02$, $p = .888$) or interaction of Test × Condition ($\chi^2(2) = 0.97$, $p = .616$) were found. Post-hoc analyses confirmed that the comprehensibility score of all the participants improved from pretest to posttest (Δ = 0.43, *SE* = 0.06, *t*(675) = 7.59, *p* < .001, *d* = 0.59) and from pretest to delayed posttest (Δ = 0.53, *SE* = 0.06, *t*(675) = 9.51, *p* < .001, *d* = 0.89). Other comparisons, however, did not reveal significant results.

Fluency. Similarly, analysis of fluency scores revealed a significant main effect of test ($\chi^2(2) = 235.91$, $p < .001$), while condition ($\chi^2(1) = 0.31$, $p = .577$) and the interaction of Test × Condition ($\chi^2(2) = 0.16$, $p = .924$) were not significant. The post-hoc analysis of test found that participants yielded continuous improvement from pretest to posttest (Δ = 0.77, *SE* = 0.06, *t*(675) = 11.93, *p* < .001, *d* = 1.12) and from posttest to delayed posttest (Δ = 0.16, *SE* = 0.06, *t*(675) = 2.42, *p* = .048, *d* = 0.23), regardless of training method.

b) Acoustic analyses

Vowel height (F1). First, although a significant main effect of test was found for /y/ ($\chi^2(2) = 9.88$, $p = .007$) and /œ/ ($\chi^2(2) = 22.54$, $p < .001$) , no significant interaction of Test × Condition was found for the F1 (Bark) in any of the three vowels (/y/: $\chi^2(2) = 0.55$, $p = .761$; /ø/: $\chi^2(2) = 1.37$, $p = .504$; /œ/: $\chi^2(2) = 4.23$, $p = .121$). This means that the F1 value did not vary across conditions. However, F1 is an indicator of vowel height (the lower the F1, the higher the vowel). The three-way contrast in height is a native feature for Catalan speakers' vocalic system. Therefore, the change in F1 is not necessarily meaningful. We thus checked whether participants could distinguish the three degrees of vowel height by further analyzing the F1 patterns. To this end, three Generalized Linear Mixed Models for F1 were additionally run with vowel (three levels: /y, ø, œ/) being the fixed effect and participants and item as random intercepts, in each of the three tests. The results revealed a significant main effect of vowel on the F1 value (Bark) in all three tests (pretest: $\chi^2(2) = 94.43$, $p < .001$; posttest: $\chi^2(2) = 92.46$, $p < .001$; delayed posttest: $\chi^2(2) = 96.79$, $p < .001$). Post-hoc analyses revealed that there was always a clear distinction in terms of F1 between the three target vowels with /y/ being lower than /ø/, while /ø/ was again lower than /œ/ in all three tests (all $p < .05$, see Table 4 for descriptive data and Table C3 for post-hoc results). This indicates that participants were able to produce the three

target vowels using three different levels of vowel height throughout the whole experiment.

Table 4

Means (Standard Deviations) of the Bark Normalized First and Second Formant Frequencies of the Three Front Rounded Vowels in the Dialogue-Reading Task Across Condition and Test

	F1 (Bark)		F2 (Bark)	
	Non-embodied	Embodied	Non-embodied	Embodied
/y/				
Pretest	4.20 (0.81)	4.15 (0.65)	11.48 (1.96)	11.30 (2.32)
Posttest	4.27 (0.79)	4.28 (0.73)	11.39 (2.07)	11.75 (1.98)
Delayed posttest	4.32 (0.72)	4.32 (0.89)	11.48 (1.80)	11.95 (1.76)
/ø/				
Pretest	4.75 (0.78)	4.88 (0.89)	11.50 (1.72)	11.40 (1.92)
Posttest	4.88 (0.78)	4.92 (0.87)	11.09 (1.81)	11.65 (1.64)
Delayed posttest	4.89 (0.82)	4.91 (0.77)	11.25 (1.62)	11.75 (1.53)
/œ/				
Pretest	5.64 (0.77)	5.78 (0.92)	11.02 (1.43)	10.89 (1.49)
Posttest	5.89 (0.74)	5.89 (0.81)	10.55 (1.58)	11.29 (1.29)
Delayed posttest	5.96 (0.89)	5.90 (0.83)	10.96 (1.33)	11.31 (1.33)

Vowel frontness/backness (F2). The second formant frequency (F2) is related to vowel frontness (the higher the F2, the more front the vowel). Given that front rounded vowels do not exist in the learners' first language (Catalan), F2 (Bark) vocalic measures were expected to improve so as to reflect a more target-like pronunciation in posttest sequences.

The analyses of all the three vowels revealed a significant interaction of Test × Condition (/y/: $\chi^2(2) = 12.00$, $p = .002$; /ø/: $\chi^2(2) = 14.38$, $p = .001$; /œ/: $\chi^2(2) = 24.46$, $p < .001$), which suggests that the change in F2 values varied across test and condition. In what follows, we report the significant changes revealed by post-hoc comparisons.

- For /y/, the F2 of the embodied group increased from pretest to posttest ($\Delta = 0.45$ Bark, $SE = 0.15$, $t(1104) = 3.12$, $p = .006$, $d = 0.33$) and from pretest to delayed posttest ($\Delta = 0.65$ Bark, $SE = 0.15$, $t(1104) = 4.52$, $p < .001$, $d = 0.47$). However, no significant difference was found for the non-embodied group in any of the three tests.

- For /ø/, the F2 of the embodied group showed a significant increase from pretest to delayed posttest ($\Delta = 0.36$ Bark, $SE = 0.14$, $t(945) = 4.52$, $p = .033$, $d = 0.29$), while the non-embodied group significantly decreased the F2 value from pretest to posttest ($\Delta = -0.42$ Bark, $SE = 0.14$, $t(945) = -3.03$, $p = .008$, $d = -0.34$).

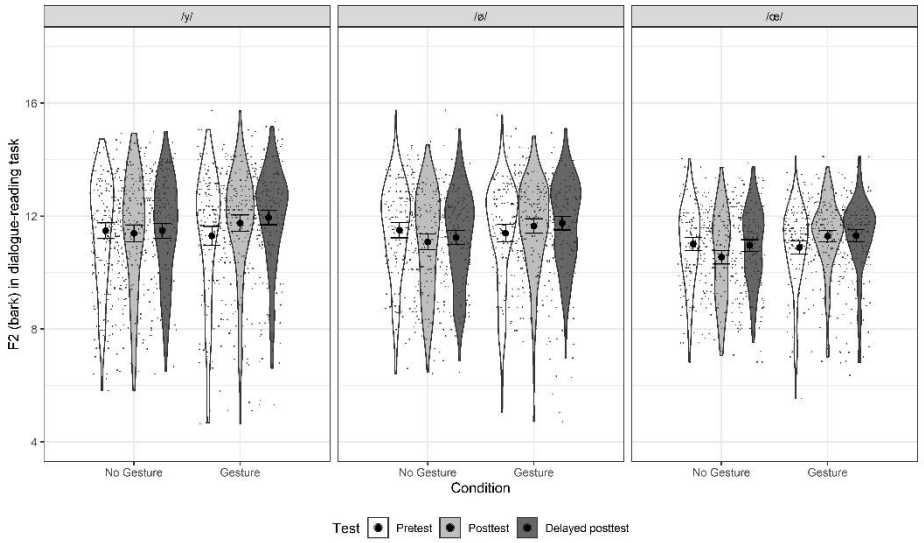
- For /œ/, the F2 of the embodied group again showed a significant increase from pretest to posttest ($\Delta = 0.40$ Bark, $SE = 0.13$, $t(945) = 3.21$, $p = .004$, $d = 0.36$) and from pretest to delayed posttest ($\Delta = 0.42$ Bark, $SE = 0.13$, $t(945) = 3.35$, p

= .003, d = 0.38). By contrast, in the non-embodied group, there was a significant decrease in F2 from pretest to posttest (Δ = -0.47 Bark, SE = 0.12, t(945) = -3.78, p = .001, d = -0.42), although this decrease was adjusted at delayed posttest by a significant improvement from posttest (Δ = 0.42 Bark, SE = 0.12, t(945) = 3.36, p = .002, d = 0.37).

Briefly, in the embodied group, all three vowels, except for /ø/, showed an increase in F2 at posttest (i.e., one week after training) and maintained this enhancement at delayed posttest (i.e., two weeks after training), which reflects a fronting effect of tongue position when producing the three target vowels. However, the non-embodied group did not show such a progressive pattern. Figure 4 visually plots the results of F2 across condition and test.

Figure 4

Mean Bark Normalized Second Formant Frequencies "F2 (Bark)" of the Three Front Rounded Vowels Across Condition and Test in the Dialogue-Reading Task



Note. The larger dots indicate the mean values. The error bars mark the 95% confidence intervals.

4.3.3 Sentence imitation task

a) Accentedness

The descriptive data for accentedness scores across condition and test are summarized in Table 5.

Table 5

Means and Standard Deviations of the Accentedness Score Across Condition and Test in the Sentence Imitation Task

	Non-embodied		Embodied	
	M	*SD*	*M*	*SD*
Pretest	4.80	1.49	4.65	1.32
Posttest	5.02	1.49	5.02	1.24

Delayed posttest	5.11	1.49	5.22	1.19

Analysis of accentedness scores in the sentence imitation task revealed a significant main effect of test ($\chi^2(2) = 104.78$, $p < .001$) as well as a significant two-way interaction of Test × Condition ($\chi^2(2) = 9.02$, $p = .011$). However, condition was not found to be significant ($\chi^2(1) = 0.00$, $p = .952$). The post-hoc analysis of the two-way interaction revealed similar patterns to those reported for the dialogue-reading task. Specifically, the non-embodied group showed a significant improvement from pretest to posttest ($\Delta = 0.22$, $SE = 0.06$, $t(2556) = 3.55$, $p = .001$, $d = 0.24$) and maintained this improvement at delayed posttest, indicated by a significant improvement from pretest to delayed posttest ($\Delta = 0.31$, $SE = 0.06$, $t(2556) = 5.08$, $p < .001$, $d = 0.34$). By contrast, the improvement observed in the embodied group was continuous, namely, a significant improvement from pretest to posttest ($\Delta = 0.37$, $SE = 0.06$, $t(2556) = 5.95$, $p < .001$, $d = 0.41$) followed by a further improvement from posttest to delayed posttest ($\Delta = 0.20$, $SE = 0.06$, $t(2556) = 3.23$, $p = .004$, $d = 0.22$).

b) Acoustic analyses

The acoustic analyses of the sentence imitation task were performed following the same procedure as that of the dialogue-reading task.

Vowel height (F1). As in the dialogue-reading task, F1 analyses found no significant two-way interaction of Test × Condition for any of the three vowels (/y/: $\chi^2(2) = 4.23$, $p = .121$; /ø/: $\chi^2(2) = 0.67$, $p = .714$; /œ/: $\chi^2(2) = 5.98$, $p = .050$). Yet a significant main effect of test was found for the three vowels (/y/: $\chi^2(2) = 10.26$, $p = .006$; /ø/: $\chi^2(2) = 13.46$, $p = .001$; /œ/: $\chi^2(2) = 12.48$, $p = .002$). In order to check if the participants distinguished the three levels of vowel height (F1 Bark), the same models were run as those applied to the dialogue-reading task. Again, for all three tests, there was a significant main effect of vowel (pretest: $\chi^2(2) = 58.74$, $p < .001$; posttest: $\chi^2(2) = 59.64$, $p < .001$; delayed posttest: $\chi^2(2) = 67.61$, $p < .001$), and post-hoc pairwise comparisons showed that the F1 value was significantly different between the three target vowels in all the three tests (all $p < .05$, see Table C4), with /y/ producing the lowest figures, and /œ/ producing the highest (see Table 6).

Table 6

Means (Standard Deviations) of the Bark Normalized First and Second Formant Frequencies of the Three Front Rounded Vowels in the Sentence Imitation Task Across Condition and Test

	F1 (Bark)		F2 (Bark)	
	Non-embodied	Embodied	Non-embodied	Embodied
/y/				
Pretest	4.08 (0.75)	4.05 (0.78)	11.69 (1.83)	11.54 (2.23)
Posttest	4.27 (0.88)	4.13 (0.77)	11.67 (1.94)	12.02 (1.57)
Delayed posttest	4.21 (0.76)	4.12 (0.74)	11.74 (1.47)	12.08 (1.43)

/ø/				
Pretest	4.71 (0.88)	4.79 (0.86)	11.32 (1.71)	11.28 (1.99)
Posttest	4.88 (0.81)	4.98 (0.99)	11.33 (1.62)	11.69 (1.45)
Delayed posttest	4.80 (0.83)	4.82 (0.80)	11.17 (1.61)	11.91 (1.38)
/œ/				
Pretest	5.69 (1.00)	5.82 (0.98)	10.80 (1.32)	10.74 (1.59)
Posttest	5.97 (0.91)	5.91 (0.80)	10.85 (1.22)	11.34 (1.07)
Delayed posttest	5.96 (0.82)	5.83 (0.70)	10.86 (1.13)	11.45 (1.06)

Vowel frontness/backness (F2). The analyses of the F2 (Bark) for all three vowels revealed a significant two-way interaction of Test × Condition (/y/: $\chi^2(2) = 7.60$, $p = .022$; /ø/: $\chi^2(2) = 15.22$, $p < .001$; /œ/: $\chi^2(2) = 19.62$, $p < .001$). This again suggests that the F2 value of the target vowels in the sentence imitation task varied across condition and test. Significant contrasts revealed by post-hoc comparisons are reported as follows.

- For /y/, the embodied group revealed a significant improvement in F2 from pretest to posttest ($\Delta = 0.48$ Bark, $SE = 0.15$, $t(1104) = 3.26$, $p = .004$, $d = 0.34$) and from pretest to delayed posttest ($\Delta = 0.54$ Bark, $SE = 0.15$, $t(1104) = 3.66$, $p = .001$, $d = 0.38$). Yet no significant contrast was observed in the non-embodied group between any of the three tests.

- For /ø/, the embodied group revealed a progressive pattern similar to that of /y/, with the pretest outperformed by the posttest ($\Delta = 0.41$ Bark, $SE = 0.14$, $t(945) = 2.88$, $p = .012$, d

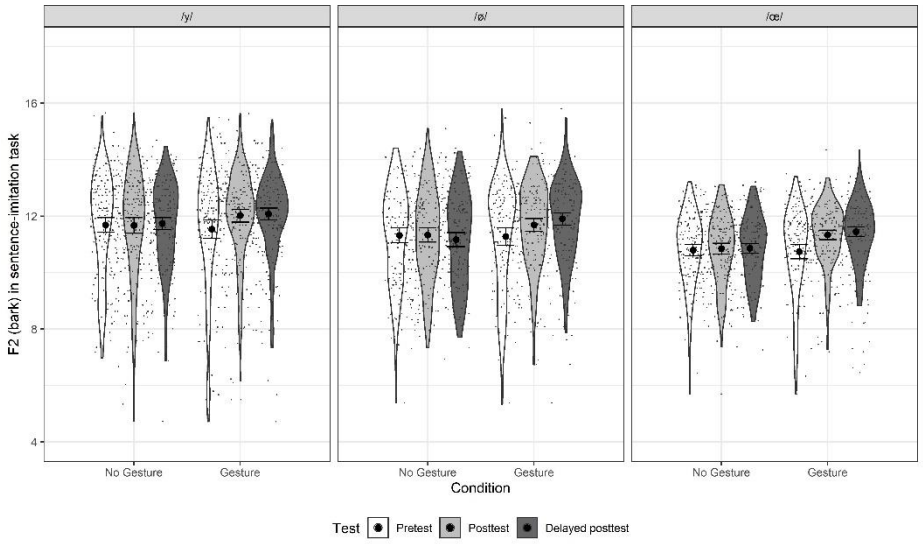
= 0.33) and delayed posttest (Δ = 0.63 Bark, *SE* = 0.14, *t*(945) = 4.43, *p* < .001, *d* = 0.50), while no significant change was shown in the non-embodied group over the three tests.

- For /œ/, again, the embodied group significantly improved their F2 from pretest to posttest (Δ = 0.59 Bark, *SE* = 0.11, *t*(945) = 5.36, *p* < .001, *d* = 0.61) and from pretest to delayed posttest (Δ = 0.71 Bark, *SE* = 0.11, *t*(945) = 6.36, *p* < .001, *d* = 0.72). The non-embodied group did not show any significant improvement or worsening between the three tests.

Summarizing, the formant analyses of the F2 component of the target front rounded vowels in the sentence imitation task revealed a similar pattern as the one reported for the dialogue-reading task. That is, while participants in the embodied group were able to front their tongue position when producing the target front rounded vowels one week after training and maintained this effect after two weeks, this was not the case for the participants in the non-embodied group. Table 6 shows the descriptive data, and Figure 5 a visual plot of the results.

Figure 5

The Bark Normalized Second Formant Frequency "F2 (Bark)" of the Three Front Rounded Vowels Across Condition and Test in the Sentence Imitation Task



Note. The larger dots indicate the mean values. The error bars mark the 95% confidence intervals.

4.4 Discussion and conclusion

The present study assessed the effects of a three-session embodied prosodic training program using a listen-and-repeat paradigm to improve both L2 pronunciation and segmental accuracy. We will organize the discussion in accordance with our two research questions.

4.4.1 Effects of embodied prosodic training on overall pronunciation proficiency

The first research question was whether embodying prosodic features through visuospatial hand gestures would improve Catalan learners'

overall French pronunciation proficiency more than parallel non-embodied training. The results for overall pronunciation proficiency showed that the pronunciation gains obtained by embodied prosodic training were larger than those achieved by non-embodied training for both reading and imitation tasks.

Regarding accentedness, the additional improvement from posttest (one week after training) to delayed posttest (two weeks after training) in both tasks in the embodied group suggested that hand gestures that highlight prosodic properties may help maintain the training effects. Therefore, our study not only supports the results of previous studies showing positive effects of embodied prosodic training techniques on pronunciation (e.g., Gluhareva & Prieto, 2017; Kushch, 2018; Llanes-Coromina et al., 2018) but also provides further evidence that gestures can play a role in the generalization and maintenance of these training effects.

By contrast, comprehensibility and fluency measures obtained in the discourse reading task did not obtain a significant two-way interaction of Test × Condition. That is, the two training methods improved the participants' performance in a similar fashion. With respect to comprehensibility, both groups significantly improved their comprehensibility score at posttest and maintained this improvement at delayed posttest with a large effect size (d = 0.89). As for fluency, the participants in both groups showed a continuous improvement across the three tests (d = 1.35).

Taken together, these results show that both embodied and non-embodied training paradigms were beneficial in improving the learners' speech comprehensibility and fluency.

Why is it the case that embodied prosodic training could not outperform non-embodied training in comprehensibility and fluency measures? In our view, this is because accentedness is mainly related to the pronunciation of segmental and suprasegmental features (e.g., K. Saito et al., 2016, 2017; Trofimovich & Isaacs, 2012), while comprehensibility is not only related to pronunciation factors but also determined by other factors like grammar, lexis (Crowther et al., 2016; K. Saito et al., 2017; Trofimovich & Isaacs, 2012), and fluency measures (Isaacs & Trofimovich, 2012). Moreover, it might well be that a three-session training program was not sufficient to improve these measures, given that in other research a longer training period (eight weeks and two hours per session) with the verbotonal method revealed a positive effect on fluency (Alazard, 2013). Future studies are needed to assess the effects of embodied training over a longer period of time.

4.4.2 Effects of embodied prosodic training on the production of non-native vocalic features

The second research question focused on whether training with hand gestures mimicking target prosodic features would benefit Catalan learners' pronunciation of non-native phonemes. Results of the formant analyses

of the three front rounded vowels /y, ø, œ/ confirmed that participants in the embodied prosodic training condition generally improved their pronunciation accuracy from pretest to posttest or to delayed posttest, in both reading and imitation tasks. This was shown by the expected F2 increase across tests, indicating a shift from back to front rounded vowels in both tasks. By contrast, participants in the non-embodied condition did not display such an improvement between any of the three tests in either of the tasks. Furthermore, in the dialogue-reading task, /ø/ and /œ/ even showed a temporary decrease (from pretest to posttest) in F2. These results indicate that a three-session embodied prosodic training program led to a more target-like pronunciation of *front* rounded vowels by helping participants move the place of articulation forward.

These findings are thus in line with the previous studies (e.g., Hardison, 2004; Missaglia, 2007; Y. Saito & Saito, 2017), showing that a focus on suprasegmental training can trigger pronunciation gains at the segmental level. The underlying mechanism may be based on several components. First, as noted above, highlighting the prosodic structure of target sentences could serve as an integrative strategy for learning the target sounds, given that they were placed in a variety of prosodic positions. Our results thus corroborate the hypotheses that there is interdependency between prosodic and segmental features in phonological learning and that prosodic features can act as scaffolding mechanisms for improving articulatory production. In addition, the visualization of target prosodic

structure through hand gestures may enhance the bootstrapping function of prosody for learning novel segmental features. This follows the Embodied Cognition paradigm, in which cognition is viewed to be strongly grounded in sensory-motor processes (Ionescu & Vasc, 2014). In relation to this, gestures have been shown to save cognitive resources and reduce the cognitive load during language processing (e.g., Cook et al., 2012; Goldin-Meadow et al., 2001). Thus, observing hand movements mimicking prosodic structure during pronunciation training may have bootstrapped phonological cognitive processes.

Moreover, participants in the embodied condition improved in terms of not only accentedness but also vocalic accuracy at posttest in both the dialogue-reading and sentence imitation tasks. However, the level of difficulty entailed by the two tasks is not the same since imitation can provide both lexical and acoustic-phonological information while only lexical pathways are available to participants in a reading task. Therefore, the consistency of these results across tasks is important because it shows that the positive effects of embodied prosodic training can be generalized from imitation tasks to more challenging reading tasks.

4.4.3 Limitations

The current study has several limitations. First, the training program applied here involved a total of three sessions, with each session lasting

only around 15 minutes. It would be of interest to assess the role of embodied prosodic training in a longitudinal study. Second, previous studies have shown that the accuracy of learners' gestural performance may have a significant impact on training effects (P. Li, Xi, et al., 2020). Since the gestures in this study were difficult to imitate in that they encoded complex prosodic information, the participants were not encouraged to produce the gestures after the instructors[6]. This design leaves it an open question as to whether learners gain more by performing gestures than by merely observing them. Finally, the two tasks in the present study were at the controlled speech level. Recent meta-analyses (e.g., K. Saito & Plonsky, 2019) suggest that spontaneous speech production should also be taken into account when assessing learners' pronunciation skills. It would thus be of interest for a future study to assess the role of gestures in spontaneous speech.

4.4.4 Conclusion

The present study showed that the use of visuospatial hand gestures mimicking prosodic features was able to not only enhance learners' overall

[6] All the videos that were filmed during training were checked to ensure that the participants followed the instructions. The results revealed only one participant tried to imitate the gestures of the instructor but after two trials, she gave up and focused on the speech repetition task. Given the fact that only two trials were affected, this participant was not removed from the analysis.

pronunciation proficiency but also improve the accuracy of specific fronting features of the front rounded vowels of Catalan learners of French. These findings help expand our understanding of the role of embodied training in L2 pronunciation and point to a possible role for discourse-based embodied pronunciation training in the L2 classrooms, where teachers could creatively use hand movements as a tool to embody the target features of L2 speech.

5

CHAPTER 5:
GENERAL DISCUSSION AND CONCLUSIONS

5.1 Summary of findings

The general goal of this PhD dissertation is to assess the potential benefits of visuospatial hand gestures encoding novel phonological properties in the context of a multimodal approach to L2 pronunciation learning. While previous studies have focused on the role of hand gestures in learning suprasegmental features or general patterns of pronunciation (e.g., Baills et al., 2019; Gluhareva & Prieto, 2017; Llanes-Coromina, Prieto, et al., 2018; Yuan et al., 2019), the present dissertation focuses on the role of embodied pronunciation training techniques for the learning of novel segmental features. We carried out three between-subject training studies with a pretest and posttest design to assess the role that gestures play in the learning of three types of non-native phonological features, namely (a) vowel-length contrasts, (b) aspiration contrasts in plosives; and (c) front rounded vowels, with each study addressing one aspect. Two of these studies also assessed the potential gains of these embodied training on global pronunciation measures, such as accentedness in both posttest and delayed posttest.

The first study investigated whether producing visuospatial hand gestures mimicking vowel length can help Catalan speakers without any knowledge of Japanese to identify and imitate long and short vowels (Chapter 2). The second study assessed the potential benefits of producing visuospatial hand gestures cueing aspirated features in learning non-

native aspirated consonants, while additionally focusing on the accuracy of participants' gesture performance (Chapter 3). Finally, the third study explored the effects of visuospatial hand gestures encoding pitch and durational properties at phrase-level, not only on the global pronunciation proficiency of French but also on the pronunciation accuracy of novel front rounded vowels (Chapter 4).

In general, the three studies jointly showed that visuospatial hand gestures encoding a variety of phonetic features (i.e., durational features, consonantal aspiration, and melodic/rhythmic features) could facilitate the L2 pronunciation at both segmental and suprasegmental levels.

The results of Study 1 showed that while participants improved equally from pretest to posttest across the two conditions (Gesture vs. No Gesture) in the identification task involving vowel-length contrasts, the Gesture group revealed more gains than the No Gesture group in the production task. These results suggest that producing hand gestures encoding durational features of speech may help novice learners acquire novel phonological contrasts in an L2.

The results of Study 2 revealed that participants who appropriately performed a fist-to-open hand gesture mimicking the air burst of the aspirated plosives during training gained enhanced voice onset time (VOT) values of the aspirated plosives. They also yielded an improvement in

the overall pronunciation of the target words, both in posttest and delayed posttest. Regarding the memorization of novel words bearing aspiration contrasts, the embodied training helped maintain word recognition accuracy after three days, while non-gestural training did not. These results suggest that producing visuospatial hand gestures encoding aspiration features can help beginner learners to more accurately produce non-native aspirated plosives. Moreover, the results emphasize that appropriate gesture performance during training is crucial in maintaining the gains obtained through the pronunciation training. Therefore, it is important to assess the learners' gesture performance in the context of embodied learning.

The results of Study 3 showed that compared to a non-embodied training, an embodied prosodic training with hand movements encoding melodic and rhythmic features of speech yielded a continuous improvement in accentedness in the dialogue-reading and sentence imitation tasks from pretest to posttest and to delayed posttest. More importantly, the embodied prosodic training improved the pronunciation accuracy of French front rounded vowels as assessed by F2 measures, which indicated that the tongue position was more fronted after training. As for comprehensibility and fluency scores, both training groups (embodied and non-embodied) showed similar levels of improvement. The results highlight the interaction between prosodic and segmental features of speech by show-

ing that training with embodied suprasegmental features has a direct beneficial effect on reducing accentedness and the acquisition of specific vocalic features.

In the following sections, we discuss the findings in relation to the beneficial role of using visuospatial hand gestures in L2 pronunciation training (section 5.2). Following this, we provide a discussion on why the use of visuospatial hand gestures helps L2 phonological learning (section 5.3). We also extend our argument regarding the benefits of combining hand gestures and the interaction between prosody and segments in L2 phonological learning (section 5.4). Finally, we close this chapter by emphasizing the theoretical and practical contributions of our findings in the field of second language pronunciation learning (section 5.5).

5.2 The effects of visuospatial hand gestures on L2 phonological learning

The underlying goal of this doctoral dissertation is to assess the role of visuospatial hand gestures in second language phonological learning from different and complementary angles. First, the three empirical studies involved a variety of target L2 features which included durational, consonantal, and vocalic articulatory features. Second, the target L2 languages included Japanese, Mandarin Chinese, and French. Third, the learning outcome was tested on different levels (from word to discourse,

and from imitation to oral reading tasks). And fourth, the learners' L2 proficiency varied from beginner to intermediate levels.

The three empirical studies revealed evidence that supports the positive role of using visuospatial hand gestures in improving learners' pronunciation accuracy of the abovementioned three features. Apart from the effects on specific segmental features, embodied pronunciation training was shown to improve the global pronunciation proficiency in an L2 (in Study 2, the general pronunciation accuracy of single words, and in Study 3, the accentedness ratings on sentences and dialogues). In other words, the three studies validated the benefits of using a set of visuospatial hand gestures for embodied L2 pronunciation training.

These findings are in line with results of previous studies showing that embodied pronunciation interventions could improve the participants' learning outcome of various phonological features, such as lexical tones (Baills et al., 2019; Morett & Chang, 2015), intonation (Kelly et al., 2017; Yuan et al., 2019), word stress (Ghaemi & Rafi, 2018), and articulatory features (Amand & Touhami, 2016; Hoetjes & van Maastricht, 2020; Xi et al., 2020); as well as global pronunciation, including measures in accentedness, comprehensibility, and fluency (Gluhareva & Prieto, 2017; Kushch, 2018; Llanes-Coromina et al., 2018). Also, the results of Study 2 added new evidence on the positive role of gestures encoding phonological features in strengthening the link between semantic meaning and

phonological forms, in line with previous studies that have proven that non-representational gestures could be of help in this domain (Baills et al., 2019; Kushch et al., 2018; Morett & Chang, 2015; So et al., 2012). Overall, these results extended the line of research from the suprasegmental domain to the segmental domain.

Another novelty of this PhD dissertation is that embodied pronunciation training revealed stronger effects at delayed posttest than non-embodied training in both pronunciation (Studies 2 and 3) and vocabulary learning (Study 2). Whereas most of the previous studies tested the training effects through an immediate posttest (Baills et al., 2019; Gluhareva & Prieto, 2017; Hirata & Kelly, 2010; Hoetjes & van Maastricht, 2020; Xi et al., 2020; Yuan et al., 2019, among many others), we tested the delayed effect over different time spans, from three days (Study 2) to two weeks (Study 3). On the one hand, for the pronunciation of consonantal and vocalic features (aspiration in Study 2 and front rounded vowels in Study 3) at delayed posttest, embodied training revealed a "maintenance" effect. That is, after gaining a significant improvement in the pronunciation accuracy of the target features from pretest to posttest, learners managed to maintain this gain at delayed posttest. On the other hand, for the general pronunciation accuracy of words (Study 2) and the accentedness of sentences and dialogues (Study 3), we observed a continuous improvement. That is, learners trained with hand gestures (Well-Performed

Gesture group in Study 2 and Embodied group in Study 3) showed significant improvements in pronunciation across pretest, posttest, and delayed posttest. As for the memorization of words bearing the target phonological contrasts (Study 2), gestural training outperformed non-gestural training as well. Taken together, the results reveal that embodying the target pronunciation features, by observing or by accurately producing hand gestures, not only improves the L2 speech production at various linguistic levels, but also has long-lasting effects compared to non-embodied training.

Moreover, in line with previous research suggesting the importance of the adequacy of gestures in training segmental features (Hoetjes & van Maastricht, 2020; Xi et al., 2020), we established that learners' gesture performance also plays a role in embodied pronunciation training. In Study 2, the speakers who did not appropriately imitate the instructor's gestures during training did not benefit from the use of these gestures. Therefore, instructors should assess whether learners are capable of accurately performing the hand gestures. A failure to appropriately imitate the instructor's gestures may be due to various reasons. First, for some participants, producing gestures while imitating non-native speech may have triggered an increase in cognitive load. That is, some tasks may be too difficult for some learners, and therefore it is the role of the instructor to propose tasks that are optimal with respect to the involvement of cog-

nitive sources. Therefore, future studies involving embodied pronunciation training may benefit from measuring learners' cognitive load (Brünken et al., 2003; van Merriënboer & Sweller, 2005) during the intervention phase. Second, it might well be that some learners lose concentration during the training, due to a lack of motivation. This variable should also be taken into consideration in future research.

Interestingly, the results of two of our studies revealed that embodied training showed limited effects on perception (Studies 1 and 2). Both durational hand gestures and hand configurations encoding aspiration features failed to help the *perception* of the target phonological features. These results are in line with previous findings that hand gestures may have limited effects on identifying non-native phonological contrasts, such as duration (e.g., Hirata et al., 2014; Hirata & Kelly, 2010) and aspiration (Xi et al., 2020). Possible reasons underlying this asymmetry are as follows. First, the perception accuracy in both Study 1 and Study 2 was already high at pretest, which suggests that the tasks were not challenging enough for participants. Future studies might want to include more challenging tasks for testing perceptual abilities. A second reason is that in both Studies 1 and 2, we asked the participants to orally produce the training words, while mimicking the hand gestures of the instructor. This may have triggered the transfer-appropriate effect, which leads to more gains in production than in perception. That is, people are more likely to recall the knowledge presented requiring the cognitive process

procedure similar to that of the training phase (Franks et al., 2000; Lightbown, 2008; Lockhart, 2002; Morris et al., 1977). Likewise, research in second language acquisition has reported that learners perform better in practiced/trained skills (M. Li & Dekeyser, 2017), suggesting that the learning outcome may be skill-specific (DeKeyser, 2015). In our case, the gestures were more tightly associated with speech production than with perception, due to the listen-and-repeat training procedure. Consequently, our embodied training may have helped participants more in production than in perception skills. Finally, as discussed in Chapters 2 and 3, the improvements in perception and production do not have to be synchronous (Nagle, 2018; Zampini, 1998). Future research may benefit from proposing multimodal training over a longer period to further examine the relationship between the improvement in the two domains.

All in all, and despite having limited effects on speech perception, the results of the present dissertation reveal consistent positive results of the hand gestures encoding novel phonetic features on L2 pronunciation.

5.3 The mechanism underlying the benefits of visuospatial hand gestures encoding phonetic features for L2 pronunciation

The positive role that producing (Studies 1 and 2) and observing (Study 3) hand gestures played in the production of non-native phonological

features can be accounted for by the Embodied Cognition theory (Barsalou, 2010; Ionescu & Vasc, 2014; Shapiro, 2014; Shapiro & Stolz, 2019, among many others). In addition, the claim that body movements can offload cognitive demands so that more cognitive resources can be used for learning (Cook et al., 2012; Goldin-Meadow, 2011; Risko & Gilbert, 2016) has been confirmed.

From the perspective of cognitive processing, instructors should seek to reduce the extraneous cognitive load (extra cognitive load caused by the instructional design) for the learners (J. Sweller et al., 1998) by using various body- and environmental-related resources. Paas and Sweller (2012) noted that working memory mainly constrains the acquisition of knowledge that humans "have not specifically evolved to acquire" (p. 29), a category to which L2 acquisition belongs. Therefore, in learning phonological knowledge of an L2, it is essential to lighten one's cognitive load so as to reduce the demands imposed on working memory. Among the suitable methods for reducing cognitive load, body movements seem to be of particular importance. Visual information grounded to body movements can be processed in an automatic and effortless manner (van Gog et al., 2009), which underscores the relevance of embodiment in education.

Under this framework, visuospatial hand gestures which encode the target phonological features are a handy tool to reduce the extraneous cognitive load (Post et al., 2013). Furthermore, it provides dynamic and embodied information (Paas & Sweller, 2012) and activates the mirror-neuron system to trigger learning by imitation (van Gog et al., 2009), which is an ability inherent to human beings (Rizzolatti & Craighero, 2004). Thus, the positive effects of visuospatial hand gestures observed in the present dissertation are likely due to the role that body movement plays in reducing the cognitive load, therefore facilitating the learning process.

A second angle which can be used to explain the mechanism underlying the positive role of gestures, is the Dual Coding theory (Clark & Paivio, 1991; Paivio, 1991). Because visual and verbal information jointly contributes to learning (Paivio, 1991, p. 260), adding visual information to the learning materials provides redundant information, which may facilitate cognitive processing mechanisms. In line with this theory, instructors' visuospatial hand gestures provide learners with visual stimuli in compensation for the verbal information. This combination results in superior learning outcomes, compared to training which only uses the verbal channel (e.g., orally providing a metalinguistic explanation of the pronunciation feature to be learned). Therefore, the results obtained in the three empirical studies, particularly the positive results of gesture ob-

servation (Study 3), add new supportive evidence to the claim that encoding information in both visual and verbal channels can reinforce learning (Clark & Paivio, 1991).

To conclude, the results of this dissertation help to consolidate the predictions of the Embodied Cognition paradigm and the Dual Coding Theory, in that they not only provide evidence for the close relationship between the body and mind but also validate the claim that the combination of visual and verbal information may reinforce the phonological learning processes. Therefore, visuospatial hand gestures, as a body movement and visual stimuli, constitute an effective tool for learning second language phonological features.

5.4 A step further: Making use of the interaction between prosodic and segmental structure for embodied pronunciation training

Following up on the positive effects of visuospatial hand gestures that explicitly encode target phonological features in an L2 (shown by Study 1 and Study 2), Study 3 went a step further in assessing the potential gains of an embodied prosodic training method. The results of the study showed that training with visuospatial hand gestures depicting phrase-level prosodic structures not only reduced the accentedness of L2 speech but also improve the pronunciation of non-native vocalic features.

First, the positive results on L2 accentedness are in line with previous research showing that training prosodic features (using both embodied and non-embodied techniques) can lead to improvements in global pronunciation proficiency, including accentedness, comprehensibility, and fluency (e.g., Derwing et al., 1998; Gluhareva & Prieto, 2017; Gordon & Darcy, 2016; Kushch, 2018; Llanes-Coromina, Prieto, et al., 2018; Missaglia, 2007; Y. Saito & Saito, 2017).

Second, the novel findings of Study 3 lie in the complex interaction between prosody, segmental features, and embodied training. As noted in Chapter 1 and discussed in Chapter 4, the interaction between prosodic and segmental (i.e., vocalic) features may affect the realization of vowels in both L1 and L2 speech (e.g., Estrada Medina, 2004, 2007; Georgeton & Fougeron, 2014; Santiago, 2021; Santiago & Mairano, 2019). The results show that embodied pronunciation training can trigger beneficial effects on the pronunciation of non-native segments (here, the front rounded vowels /y, ø, œ/). Adding gestures to depict the prosodic structures of the training sentences (i.e., the embodied prosodic training paradigm) can thus reinforce the interaction between prosody and vowels. Because hand gestures here serve as a visualizer and an embodied form of the phrase-level prosodic structure, their function is two-fold. On the one hand, they provide visual information to speech and activate both verbal and visual channels so that the redundant information helps learners to process the learning materials. On the other hand, embodying the

prosodic structure through the use of hand gestures reinforces the phonological information for the learners (Cook et al., 2012; Paas & Sweller, 2012; Risko & Gilbert, 2016) so that it further promotes learning.

Taken together, the results of Study 3 reveal a complex picture of the interaction between prosody, segments, and the use of hand gestures. It does this by showing that embodied prosodic training involving gestures that highlight phrase-level prosodic structure may trigger positive results in the pronunciation of specific segmental features.

5.5 Final remarks: Implications, limitations, and conclusion

This doctoral dissertation focused on the beneficial effects of hand gestures in L2 phonological learning. By conducting three empirical studies, we have demonstrated that visuospatial hand gestures encoding a variety of segmental and suprasegmental features can have beneficial effects on L2 pronunciation learning. Thus, the three studies fall under the framework of the Embodied Cognition paradigm, which proposes that body and mind are closed related to each other. Consequently, the present dissertation has the following theoretical and practical implications.

5.5.1 Implications for the Embodied Cognition paradigm

As has been noted in Chapter 1, the three studies were proposed under the framework of Embodied Cognition (e.g., Barsalou, 2008; Foglia & Wilson, 2013; Ionescu & Vasc, 2014; Kiefer & Trumpp, 2012; Kontra et al., 2015; Mizelle & Wheaton, 2010; Shapiro, 2011; Shapiro & Stolz, 2019; Wilson, 2002). The positive results on visuospatial hand gestures confirm the predictions of the Embodied Cognition paradigm as applied to phonological learning.

Embodied Cognition paradigm holds that there is a tight connection between manual movements and speech motor actions (Gentilucci, 2003; Gentilucci & Corballis, 2006) and that the coordination and interaction between speech and gestures stem from the entrainment of the two motor systems (Rusiewicz et al., 2013, 2014). The results of the present dissertation thus provide new evidence for the connection between speech and gestures in the context of L2 pronunciation learning. On the one hand, embodied training paradigm can trigger beneficial effects on L2 pronunciation learning, which is in line with the claim that body movement is a way of shifting cognitive load (Risko & Gilbert, 2016). However, on the other hand, the gesture performance quality of the learners influences their learning outcome (i.e., learners benefit from appropriately performing the target gesture more than not doing so), which conforms to the claim that our body partially constrains our cognition (Shapiro, 2011;

Wilson, 2002). All in all, the results of this dissertation back up the theory of the Embodied Cognition paradigm and extend its implications to the learning of L2 phonology, particularly at the segmental level.

5.5.2 Implications for the multimodal or multisensory approaches to L2 pronunciation teaching

Our results have added new evidence in favor of the multimodal and multisensory approaches to L2 pronunciation teaching.

First, Odisho (2007, 2014) proposed the Multisensory, Multicognitive Approach to L2 learning, which encourages the integration of different sensory *modalities*, especially the visual and tactile-kinesthetic ones, instead of the traditional exclusive sensory modality (i.e., auditory sensory modality only) when instructing L2 pronunciation. According to Odisho (2007), the pronunciation teaching procedure integrates the following orientations: cognitive orientation, auditory orientation, visual orientation, and kinesthetic/proprioceptive orientation. Since instructors and learners should both be involved in teaching activities in a multisensory fashion, the traditional teaching methods that rely on auditory modalities should no longer be encouraged. Furthermore, it is also promoted that teachers should make use of the body and facial gestures, which are "extremely helpful in teaching pronunciation." These benefits range from the learning of phonemic features to suprasegmental features (Odisho,

2014, p. 81). As embodied training with hand gestures involves multimodal and multisensory input, the three training studies in this PhD dissertation fall in the scope of multimodal or multisensory training paradigm. Therefore, the positive results obtained from gestural training support the claim that language training, especially pronunciation training, should be multimodal and multisensory.

Second, verbotonal method (Guberina, 2008; Intravaia, 2000; Renard, 1979) also involves multimodal/multisensory training techniques. This method promotes the use of prosodic structures to perform phonetic corrections in L2 pronunciation instruction with hand movements as an aiding tool. The present dissertation is in line with this training method. The results of Study 3 provide positive evidence in the belief that illustrating prosodic features may have positive effects on improving segmental accuracy in an L2 (Billières, 2002; Renard, 2002).

Moreover, according to the Contrastive Prosody Method (Missaglia, 2007), phonetic corrections on the segmental level should be done alongside the corrections on the prosodic level. Our results from Study 3 validated this claim by showing that highlighting prosodic structures with hand gestures resulted in gains in segmental accuracy (i.e., the front rounded /y, ø, œ/).

To summarize, as it has been suggested that L2 pronunciation training should be done in a multimodal/multisensory fashion, the present dissertation contributes to this field. It provides three multimodal training studies with positive evidence to support the effectiveness of these training proposals. However, more empirical work is needed to test further their practical value from various angles. We will expand this point in Section 5.5.4.

5.5.3 Implications for L2 teaching pronunciation practice

Our results provide further empirical evidence on the value of hand gestures encoding phonological features in classroom teaching practice. While a number of classroom observations (Hudson, 2011; Rosborough, 2010) and teaching proposals (Chan, 2018; Roberge et al., 1996; Smotrova, 2017; Y. Zhang, 2002) have already promoted their use in L2 classrooms, it is not until recently that researchers have started to empirically assess the role they play in learning L2 pronunciation.

As for the implications for teaching practice, the present dissertation brings the following points to the table.

First, instructors can be creative in designing and using visuospatial hand gestures. Hand gestures are a useful tool as they do not require any specific technology and can be easily imitated by the learners. The gesture

shapes used in this dissertation were created based on neuroscientific research or borrowed from previous teaching suggestions. In practice, instructors can create their own gestures based on their teaching needs.

Second, the shape of visuospatial hand gestures should be appropriately designed and performed. As is shown by Study 2 and related studies (P. Li, Xi, et al., 2020; Xi et al., 2020), if gestures misrepresent the target feature to be learned, or if learners cannot manage to imitate the gestures appropriately, even if they are well designed, multimodal training may reveal null results. Thus, in teaching practice, instructors should be cautious about the learners' gesture performance during training.

Third, visuospatial hand gestures can be an interactive communication tool in classroom teaching. As discussed in Study 2, the quality of gesture performance might be an indication of the learners' cognitive load, learning motivation, effort, and so on. Although visuospatial hand gestures do not convey semantic meaning, by carefully observing the gestures produced by learners, instructors can be aware of their learning status and try to adjust the instruction accordingly.

To sum up, both instructors and learners are encouraged to make use of hand gestures. As pointed out by Shapiro and Stolz (2019), "embodiment offers either a causal route to more effective learning or a diagnostic tool for measuring conceptual understanding" (p. 30).

5.5.4 Limitations and potential future studies

First, the studies included in the present dissertation assessed the role of a handful of hand configurations and movements that supported the learning of vowel-length contrasts, consonantal aspiration, and the fronting feature of vowels. Future studies could test the role of embodiment on the learning of other relevant features like rhotacization of vowels (e.g., the /ɚ/ in "teacher" and in the so-called "erization" of Mandarin syllable finals), retroflex consonants (e.g., /ʂ/ vs. /s/ in Mandarin), vowel openness (e.g., the /e-ɛ/ and /o-ɔ/ contrasts in Catalan), pharyngealization of consonants (e.g., the /d-dˤ/ contrast in Arabic), the tense-lax vowel contrasts (e.g., /i/ vs. /ɪ/ in Englis), and so forth. By expanding the current line of research to a larger set of segmental features, a complete picture of an embodied pronunciation paradigm could emerge and be empirically tested within second language research.

Second, the effectiveness of embodied pronunciation training should be comprehensively assessed over a longer course of training, including various measures of phonological knowledge and involving learners with different levels of proficiency. Although a three-week training study was conducted for Study 3, which included both posttest and delayed posttest, it would have been interesting to conduct a longitudinal study over a semester or an academic year, to observe the L2 phonological development under the embodied training paradigm.

Third, the results of our training studies showed an asymmetry between L2 speech perception and production. While the positive role of hand gestures on production patterns was validated, this was not the case for perception patterns. In some domains there are still mixed results, such as the asymmetric effects of hand gestures on perception and production (e.g., Hirata et al., 2014; Hirata & Kelly, 2010; P. Li, Baills, et al., 2020; Xi et al., 2020), as well as some null effects in speech production (e.g., van Maastricht et al., 2019; Zheng et al., 2018). Even though the development of perception and production may not be synchronous, more research is still needed to assess their relationship.

Fourth, L2 pronunciation in the present dissertation is measured by word imitation, sentence imitation, and dialogue reading tasks, but no spontaneous and communicatively relevant tasks are included. Future studies might want to test the role of embodied pronunciation training in spontaneous speech with spontaneous production tasks, such as the TPD task (Timed Picture-Description task, see K. Saito & Munro, 2014), the picture description task with loaded sentences containing the target features to be investigated (R. Zhang & Yuan, 2020), or semi-structured oral interviews in the target L2 (He, 2014; He et al., 2015).

Fifth, the current dissertation did not directly compare the effects of training with gesture observation and gesture production in L2 pronunciation learning. Previous research did not show consistent and sufficient

evidence in this regard. For example, producing and observing hand gestures were found to have similar effects on the perception of duration (Hirata et al., 2014; Kelly et al., 2017) and lexical tones (Baills et al., 2019) by beginner learners, it is not clear whether the same holds true for speech production, especially on segmental level, by learners with more advanced proficiencies. Future studies therefore might want to make comparisons between the two training methods on the segmental learning in an L2 with various proficiency levels.

Finally, individual differences should receive more attention. It has been well documented that individual differences may account for learners' second language speech (Rota & Reiterer, 2009). Especially musicality (see Chobert & Besson, 2013 for a review) and working memory (Baddeley, 2003; Wen, 2018). In Studies 1-2, we have controlled for these two aspects by a perceptual musical test (Study 1), a musical background questionnaire (Studies 1-2), and a working memory test (Studies 1-2). However, the interaction of embodied training and individual cognitive aspects still remains to be explored. Moreover, it appears that other cognitive measures should be investigated as well, such as learning motivation and personality (e.g., Rizvanović, 2018), phonetic/speech imitation abilities (e.g., Reiterer, 2019), and so on. It is therefore interesting for future studies to assess how these individual differences interact with embodiment in L2 pronunciation training.

In conclusion, although with certain limitations, the present PhD dissertation offers consistent results from three training studies which show that visuospatial hand gestures play a positive role in L2 pronunciation learning, especially on the acquisition of novel segmental features. Thus, the results do not only provide further evidence for adopting an embodied training paradigm for the learning of second language pronunciation, but they also offer new evidence to expand our understanding of the relationship between the body and the mind, and its role in learning.

BIBLIOGRAPHY

Alazard-Guiu, C., Santiago, F., & Mairano, P. (2018). L'incidence de la correction phoné- tique sur l'acquisition des voyelles en langue étrangère : étude de cas d'anglophones apprenant le français. *Journées d'Études Sur La Parole, JEP 2018*, 116–124. https://doi.org/10.21437/jep.2018-14 . hal-01779110 HAL

Alazard, C. (2013). *Rôle de la prosodie dans la fluence en lecture oralisée chez des apprenants de Français Langue Étrangère* [Université Toulouse le Mirail]. https://doi.org/10.13140/RG.2.2.24924.23686

Alazard, C., Astésano, C., & Billières, M. (2010). The implicit prosody hypothesis applied to foreign language learning: From oral abilities to reading skills. *Proceedings of the 5th International Conference on Speech Prosody*, 1–4.

Aliaga-Garcia, C., Mora, J. C., & Cerviño-Povedano, E. (2010). L2 speech learning in adulthood and phonological short-term memory. In K. Dziubalska-Kołaczyk, M. Wrembel, & M. Kul (Eds.), *Proceedings of the 6th International Symposium on the Acquisition of Second Language Speech, New Sounds 2010* (Vol. 47, Issue 1, pp. 1–14). https://doi.org/10.2478/psicl-2011-0002

Allen, L. Q. (1995). The effects of emblematic gestures on the development and access of mental representations of French expressions. *The Modern Language Journal*, *79*(4), 521–529. https://doi.org/10.1111/j.1540-4781.1995.tb05454.x

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

Anderson-Hsieh, J., Johnson, R., & Koehler, K. (1992). The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure. *Language Learning*, *42*(4), 529–555. https://doi.org/10.1111/j.1467-1770.1992.tb01043.x

Baddeley, A. (2003). Working memory and language: An overview. *Journal of Communication Disorders*, *36*(3), 189–208. https://doi.org/10.1016/S0021-9924(03)00019-4

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. https://doi.org/10.1016/j.jml.2015.10.008

Baills, F., & Prieto, P. (2021). Embodying rhythmic properties of a foreign language through hand-clapping helps children to better pronounce words. *Language Teaching Research*, 1–31. https://doi.org/10.1177/1362168820986716

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

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.), *9th International Conference on Speech Prosody 2018* (pp. 853–857). 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 Research Forum* (pp. 187–198).

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

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

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear Mixed-Effects Models using {lme4}. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Beckman, M., Edwards, J., & Fletcher, J. (1992). Prosodic structure and tempo in a sonority model of articulatory dynamics. In G. J. Docherty & D. R. Ladd (Eds.), *Papers in Laboratory Phonology II: Gesture, Segment, Prosody* (pp. 68–89). https://doi.org/10.1017/cbo9780511519918.004

Bedore, L. M., & Leonard, L. B. (1995). Prosodic and syntactic bootstrapping and their clinical applications. *American Journal of Speech-Language Pathology*, *4*(1), 66–72. https://doi.org/10.1044/1058-0360.0401.66

Billières, M. (2002). Le corps en phonétique corrective. In R. Renard (Ed.), *Apprentissage d'une Langue Étrangère/Seconde Vol.2 La Phonétique Verbo-tonale* (pp. 38–70). De Boeck Université.

Boersma, P., & Weenink, D. (2017). *Praat: doing phonetics by computer [computer program]* (6.0.28). http://www.praat.org

Boersma, P., & Weenink, D. (2020). *Praat: doing phonetics by computer [computer program]* (6.1.16). 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), 1–16. https://doi.org/10.5334/labphon.91

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*, *19*(5), 971–986. https://doi.org/10.1017/S1366728915000425

Borghi, A. M., & Caruana, F. (2015). Embodiment theory. In J. D. Wright (Ed.), *International Encyclopedia of the Social & Behavioral Sciences* (2nd ed., Vol. 7, pp. 420–426). Elsevier. https://doi.org/10.1016/B978-0-08-097086-8.56025-5

Boucher, V. J. (2002). Timing relations in speech and the identification of voice-onset times: A stable perceptual boundary for voicing categories across speaking rates. *Perception and Psychophysics*, *64*(1), 121–130. https://doi.org/10.3758/BF03194561

Boula de Mareüil, P., & Vieru-Dimulescu, B. (2006). The contribution of prosody to the perception of foreign accent. *Phonetica*, *63*(4), 247–267. https://doi.org/10.1159/000097308

Bowles, A. R., Chang, C. B., & Karuzis, V. P. (2016). Pitch Ability As an Aptitude for Tone Learning. *Language Learning*, *66*(4), 774–808. https://doi.org/10.1111/lang.12159

Broaders, S. C., Cook, S. W., Mitchell, Z., & Goldin-Meadow, S. (2007). Making Children Gesture Brings Out Implicit Knowledge and Leads to Learning. *Journal of Experimental Psychology:*

General, *136*(4), 539–550. https://doi.org/10.1037/0096-3445.136.4.539

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Machler, M., & Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, *9*(2), 378–400.

Brünken, R., Plass, J. L., & Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning. *Educational Psychologist*, *38*(1), 53–61. https://doi.org/10.1207/S15326985EP3801_7

Byrd, C. E., Mcneil, N. M., Mello, S. K. D., & Cook, S. W. (2014). Gesturing May Not Always Make Learning Last. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the Annual Meeting of the Cognitive Science Society* (pp. 1982–1987). Cognitive Science Society.

Cai, Z. G., & Connell, L. (2012). Space-time interdependence and sensory modalities: Time affects space in the hand but not in the eye. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 168–173). Cognitive Science Society.

Cai, Z. G., Connell, L., & Holler, J. (2013). Time does not flow without language: Spatial distance affects temporal duration regardless of movement or direction. *Psychonomic Bulletin and Review*, *20*(5), 973–980. https://doi.org/10.3758/s13423-013-0414-3

Campfield, D. E., & Murphy, V. A. (2014). Elicited imitation in search of the influence of linguistic rhythm on child L2 acquisition. *System*, *42*(1), 207–219. https://doi.org/10.1016/j.system.2013.12.002

Casasanto, D., & Boroditsky, L. (2008). Time in the mind: Using space to think about time. *Cognition*, *106*(2), 579–593. https://doi.org/10.1016/j.cognition.2007.03.004

Casasanto, D., Phillips, W., & Boroditsky, L. (2003). Do we think about music in terms of space? Metaphoric representation of musical pitch. In R. Alterman & D. Kirsh (Eds.), *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 25, p. 1323). Cognitive Science Society.

Chan, M. J. (2018). Embodied pronunciation learning: Research and practice. *The Catesol Journal*, *1*, 47–68. https://files.eric.ed.gov/fulltext/EJ1174234.pdf

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction*, *8*(4), 293–332. https://doi.org/10.1207/s1532690xci0804_2

Chao, K., & Chen, L. (2008). A Cross-linguistic study of Voice Onset Time in stop consonant productions. *Computational Linguistics and Chinese Language Processing*, *13*(2), 215–232.

Chen, C.-M. (2013). Gestures as tone markers in multilingual communication. In I. Kecskes (Ed.), *Research in Chinese as a Second Language* (pp. 143–168). Walter De Gruyter. https://doi.org/10.1515/9781614512554.143

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. In F. Bimbot, C. Cerisara, C. Fougeron, G. Gravier, L. Lamel, F. Pellegrino, & P. Perrier (Eds.), *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* (Issue August, pp. 2370–2374).

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

Christophe, A., Guasti, T., & Nespor, M. (1997). Reflections on Phonological Bootstrapping Its Role.pdf. *Language and Cognitive Processes*, *12*(5), 585–612.

Christophe, A., Millotte, S., Bernal, S., & Lidz, J. (2008). Bootstrapping lexical and syntactic acquisition. *Language and Speech*, *51*(1–2), 61–75. https://doi.org/10.1177/00238309080510010501

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

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Lawrence Erlbaum Associates.

Connell, L., Cai, Z. G., & Holler, J. (2013). Do you see what I'm singing? Visuospatial movement biases pitch perception. *Brain and Cognition*, *81*, 124–130.

Cook, S. W., Yip, T. K. Y., & Goldin-Meadow, S. (2012). Gestures, but not meaningless movements, lighten working memory load when explaining math. *Language and Cognitive Processes*, *27*(4), 594–610. https://doi.org/10.1080/01690965.2011.567074

Cooper, A., & Wang, Y. (2012). The influence of linguistic and musical experience on Cantonese word learning. *The Journal of the Acoustical Society of America*, *131*(6), 4756–4769. https://doi.org/10.1121/1.4714355

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching and Assessment*. Cambridge University Press.

Crowther, D., Kim, S., Lee, J., Lim, J., & Loewen, S. (2021). Methodological synthesis of cluster analysis in second language research. *Language Learning*, *71*(1), 99–130. https://doi.org/10.1111/lang.12428

Crowther, D., Trofimovich, P., & Isaacs, T. (2016). Linguistic dimensions of second language accent and comprehensibility. *Journal of Second Language Pronunciation*, *2*(2), 160–182. https://doi.org/10.1075/jslp.2.2.02cro

Dankovičová, J., House, J., Crooks, A., & Jones, K. (2007). The relationship between musical skills, music training, and intonation analysis skills. *Language and Speech*, *50*(2), 177–225. https://doi.org/10.1177/00238309070500020201

Darcy, I., Dekydtspotter, L., Sprouse, R. A., Glover, J., Kaden, C., McGuire, M., & Scott, J. H. G. (2012). Direct mapping of acoustics to phonology: On the lexical encoding of front rounded vowels in L1 English- L2 French acquisition. In *Second Language Research* (Vol. 28, Issue 1). https://doi.org/10.1177/0267658311423455

de Bot, K. (1983). Visual feedback of intonation I: Effectiveness and induced practice behavior. *Language and Speech*, *26*(4), 331–350.

de Jong, K. J. (1995). The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *The Journal of Acoustic Society of America*, *97*(1), 491–504.

de Ruiter, J. P. (2017). The asymmetric redundancy of gesture and speech. In R. B. Church, M. W. Alibali, & S. D. Kelly (Eds.), *Why*

Gesture?: How the Hands Function in Speaking, Thinking and Communicating (pp. 59–75). John Benjamin's Publishing Company.

de Ruiter, J. P., Bangerter, A., & Dings, P. (2012). The interplay between gesture and speech in the production of referring expressions: Investigating the tradeoff hypothesis. *Topics in Cognitive Science*, *4*(2), 232–248. https://doi.org/10.1111/j.1756-8765.2012.01183.x

DeKeyser, R. (2015). Skill Acquisition Theory. In B. VanPatten & J. Williams (Eds.), *Theories in Second Language Acquisition: An Introduction* (pp. 94–112). Routledge.

Delais-Roussarie, E., Post, B., Avanzi, M., Buthke, C., Di Cristo, A., Feldhausen, I., Jun, S.-A., Martin, P., Meisenburg, T., Rialland, A., Sichel-Bazin, R., & Yoo, H.-Y. (2015). Intonational phonology of French: Developing a ToBI system for French. In S. Frota & P. Prieto (Eds.), *Intonation in Romance* (pp. 63–100). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199685332.003.0003

Derwing, T. M., Munro, M. J., & Wiebe, G. (1998). Evidence in favor of a broad framework for pronunciation instruction. *Language Learning*, *48*(3), 393–410. https://doi.org/10.1111/0023-8333.00047

Desai, R. H., Binder, J. R., Conant, L. L., & Seidenberg, M. S. (2010). Activation of sensory-motor areas in sentence comprehension. *Cerebral Cortex*, *20*(2), 468–478. https://doi.org/10.1093/cercor/bhp115

Dolscheid, S., Willems, R. M., Hagoort, P., & Casasanto, D. (2014). The relation of space and musical pitch in the brain. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *The 36th*

Annual Meeting of the Cognitive Science Society (Vol. 3, Issue 2014, pp. 421–426).

Duanmu, S. (2007). *The Phonology of Standard Mandarin*. Oxford University Press.

Estrada Medina, M. (2004). L'expression de l'emotion et la correction phonetique: L'exemple de la surprise. In R. López Carrillo & J. Suso López (Eds.), *Le Français Face aux Défis Actuels: Histoire, Langue et Culture* (pp. 319–329). Editorial Universidad de Granada.

Estrada Medina, M. (2007). Incidence de la prosodie sur la structuration de la matiere phonique : L'exemple de la surprise. *Actes Du Deuxième Colloque International de Didactique Cognitive Des Langues*, 87–94.

Foglia, L., & Wilson, R. A. (2013). Embodied cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*(3), 319–325. https://doi.org/10.1002/wcs.1226

Fortkamp, M. B. M. (2000). *Working Memory Capacity and L2 Speech Production: An Exploratory Study*. Universidade Federal de Santa Catarina,Brazil.

Fougeron, C., & Jun, S.-A. (2002). Realizations of accentual phrase in French intonation. *Probus*, *14*(2002), 147–172.

Fox, J., & Weisberg, S. (2019). *An {R} Companion to Applied Regression* (Third). Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Franks, J. J., Bilbrey, C. W., Khoo Guat Lien, & McNamara, T. P. (2000). Transfer-appropriate processing (TAP) and repetition

priming. *Memory and Cognition*, *28*(7), 1140–1151. https://doi.org/10.3758/BF03211815

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2019). *irr: Various Coefficients of Interrater Reliability and Agreement version 0.84.1[software]*. https://cran.r-project.org/package=irr

Gentilucci, M. (2003). Grasp observation influences speech production. *European Journal of Neuroscience*, *17*(1), 179–184. https://doi.org/10.1046/j.1460-9568.2003.02438.x

Gentilucci, M., & Corballis, M. C. (2006). From manual gesture to speech: A gradual transition. *Neuroscience and Biobehavioral Reviews*, *30*(7), 949–960. https://doi.org/10.1016/j.neubiorev.2006.02.004

Georgeton, L., & Fougeron, C. (2014). Domain-initial strengthening on French vowels and phonological contrasts: Evidence from lip articulation and spectral variation. *Journal of Phonetics*, *44*(1), 83–95. https://doi.org/10.1016/j.wocn.2014.02.006

Ghaemi, F., & Rafi, F. (2018). The impact of visual aids on the retention of English word stress patterns. *International Journal of Applied Linguistics and English Literature*, *7*(2), 225. https://doi.org/10.7575/aiac.ijalel.v.7n.2p.225

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

Goldin-Meadow, S. (2010). When gesture does and does not promote learning. *Language and Cognition*, *2*(1), 1–19. https://doi.org/10.1515/langcog.2010.001

Goldin-Meadow, S. (2011). Learning through gesture. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(6), 595–607. https://doi.org/10.1002/wcs.132

Goldin-Meadow, S. (2018). Taking a Hands-on Approach to Learning. *Policy Insights from the Behavioral and Brain Sciences*, *5*(2), 163–170. https://doi.org/10.1177/2372732218785393

Goldin-Meadow, S., Cook, S. W., & Mitchell, Z. A. (2009). Gesturing gives children new ideas about math. *Psychological Science*, *20*(3), 267–272. https://doi.org/10.1111/j.1467-9280.2009.02297.x

Goldin-Meadow, S., Nusbaum, H., Kelly, S. D., & Wagner, S. (2001). Explaining Math: Gesturing Lightens the Load. *Psychological Science*, *12*(6), 516–522. https://doi.org/10.1111/1467-9280.00395

Goldin-Meadow, S., & Wagner, S. M. (2005). How our hands help us learn. *Trends in Cognitive Sciences*, *9*(5), 234–241. https://doi.org/10.1016/j.tics.2005.03.006

Gordon, J., & Darcy, I. (2016). The development of comprehensible speech in L2 learners. *Journal of Second Language Pronunciation*, *2*(1), 56–92. https://doi.org/10.1075/jslp.2.1.03gor

Gordon, J., Darcy, I., & Ewert, D. (2013). Pronunciation teaching and learning: Effects of explicit phonetic instruction in the L2 classroom. In J. Levis & K. LeVell (Eds.), *Proceedings of the 4th Pronunciation in Second Language Learning and Teaching Conference* (pp. 194–206). Iowa State University.

Gorjian, B., Hayati, A., & Pourkhoni, P. (2013). Using Praat Software in Teaching Prosodic Features to EFL Learners. *Procedia - Social and Behavioral Sciences*, *84*, 34–40. https://doi.org/10.1016/j.sbspro.2013.06.505

Guberina, P. (2008). *Retrospección* (J. Murillo (ed. & trans.)). Éditions du CIPA - Asociación Española Verbotonal.

Gullberg, M. (2006). Some reasons for studying gesture and second language acquisition (Hommage à Adam Kendon). *International Review of Applied Linguistics in Language Teaching*, *44*(2), 103–124. https://doi.org/10.1515/IRAL.2006.004

Gullberg, M. (2014). Gestures and second language acquisition. In C. Müller, A. Cienki, E. Fricke, S. H. Ladewig, D. McNeill, & S. Tessendorf (Eds.), *Body-Language- Communication: An International Handbook on Multimodality in Human Interaction* (Issue 2, pp. 1868–1875). Mouton de Gruyter.

Hallgren, K. A. (2012). Computing Inter-Rater Reliability for Observational Data: An Overview and Tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 23–34. https://doi.org/10.20982/tqmp.08.1.p023

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*(DEC), 1–15. https://doi.org/10.3389/fpsyg.2017.02051

Hannahs, S. J. (2007). French phonology and L2 acquisition. In A. Dalila (Ed.), *French Applied Linguistics* (pp. 50–74). John Benjamins.

Hardison, D. M. (2004). Generalization of computer-assisted prosody training: Quantitative and qualitative findings. *Language Learning & Technology*, *8*(1), 34–52. http://llt.msu.edu/vol8num1/hardison/

Hay, J. F., Sato, M., Coren, A. E., Moran, C. L., & Diehl, R. L. (2006). Enhanced contrast for vowels in utterance focus: A cross-language

study. *The Journal of the Acoustical Society of America*, *119*(5), 3022–3033. https://doi.org/10.1121/1.2184226

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*. https://doi.org/10.1016/j.specom.2005.04.007

He, B. (2014). *Improving the English pronunciation of Chinese EFL learneres through the integration of CALL and verbotonalism* [Suranaree University of Technology]. https://doi.org/http://sutir.sut.ac.th:8080/jspui/handle/123456789/5370

He, B., Sangarun, P., & Lian, A. (2015). Improving the English pronunciation of Chinese EFL university students through the integration of CALL and verbotonalism. In J. Colpaert, A. Aerts, M. Oberhofer, & M. Gutiérez-Colón Plana (Eds.), *Seventeenth International CALL Research Conference: Task Design and CALL* (pp. 276–285). Universiteit Antwerpen.

Herrick, D. (2003). *An acoustic analysis of phonological vowel reduction in six varieties of Catalan*. University of California, Santa Cruz.

Hincks, R., & Edlund, J. (2009). Promoting increased pitch variation in oral presentations with transient visual feedback. *Language Learning and Technology*, *13*(3), 32–50.

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

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

Hostetter, A. B., & Alibali, M. W. (2008). Visible embodiment: Gestures as simulated action. *Psychonomic Bulletin and Review*, *15*(3), 495–514. https://doi.org/10.3758/PBR.15.3.495

Huang, X., Kim, N., & Christianson, K. (2019). Gesture and Vocabulary Learning in a Second Language. *Language Learning*, *69*(1), 177–197. https://doi.org/10.1111/lang.12326

Hudson, N. (2011). *Teacher gesture in a post-secondary English as a second language classroom: A sociocultural approach*. University of Nevada Las Vegas.

IBM Cooperation. (2016). *IBM SPSS statistics for Windows (version 24.0)*. IBM Cooperation.

Intravaia, P. (2000). *Formation des Professeurs de Langue en Phonétique Corrective.Le Système Verbo-tonal*. Didier Érudition.

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

Isaacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility. *Studies in Second Language Acquisition, 34*(3), 475–505. https://doi.org/10.1017/S0272263112000150

Iverson, J. M., Capirci, O., Longobardi, E., & Caselli, C. M. (1999). Gesturing in mother-child interactions. *Cognitive Development, 14*, 57–75.

Iverson, J. M., & Goldin-Meadow, S. (2005). Gesture paves the way for language development. *Psychological Science, 16*(5), 367–371. https://doi.org/10.1111/j.0956-7976.2005.01542.x

Johnson, K. (2011). *Acoustic and Auditory Phonetics* (3rd ed.). Blackwell Publishing.

Juffs, A., & Harrington, M. (2011). Aspects of working memory in L2 learning. *Language Teaching*, *44*(2), 137–166. https://doi.org/10.1017/S0261444810000509

Jun, S.-A., & Fougeron, C. (2000). A phonological model of French intonation. In A. Botinis (Ed.), *Intonation: Analysis, Modeling and Technology* (pp. 209–242). Kluwer Academic Publishers. https://doi.org/10.1007/978-94-011-4317-2_10

Kang, E. Y., Sok, S., & Han, Z. H. (2019). Thirty-five years of ISLA on form-focused instruction: A meta-analysis. *Language Teaching Research*, *23*(4), 428–453. https://doi.org/10.1177/1362168818776671

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), 1–11. https://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, G. Pinter, S. Ogawa, M. Giriko, & H. Takeyasu (Eds.), *New Development in Phonology Research: Festschrift in Honor of Haruo Kubozono* (pp. 278–294). Kaitakusha.

Kelly, S. D., Hirata, Y., Manansala, M., & Huang, J. (2014). Exploring the role of hand gestures in learning novel phoneme contrasts and vocabulary in a second language. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2014.00673

Kelly, S. D., & Lee, A. L. (2012). When actions speak too much louder than words: Hand gestures disrupt word learning when phonetic demands are high. *Language and Cognitive Processes*, *27*(6), 793–807. https://doi.org/10.1080/01690965.2011.581125

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

Kiefer, M., & Trumpp, N. M. (2012). Embodiment theory and education: The foundations of cognition in perception and action. *Trends in Neuroscience and Education*, *1*(1), 15–20. https://doi.org/10.1016/j.tine.2012.07.002

Kita, S., Alibali, M. W., & Chu, M. (2017). How do gestures influence thinking and speaking? The gesture-for-conceptualization hypothesis. *Psychological Review*, *124*(3), 245–266. https://doi.org/10.1037/rev0000059

Kita, S., & Özyürek, A. (2003). What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking.

Journal of Memory and Language, *48*(1), 16–32. https://doi.org/10.1016/S0749-596X(02)00505-3

Kleber, F. (2018). VOT or quantity: What matters more for the voicing contrast in German regional varieties? Results from apparent-time analyses. *Journal of Phonetics*, *71*, 468–486. https://doi.org/10.1016/j.wocn.2018.10.004

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

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Krahmer, E., & Swerts, M. (2007). The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language*, *57*(3), 396–414. https://doi.org/10.1016/j.jml.2007.06.005

Krauss, R. M. (1998). Why do we gesture when we speak? *Current Directions in Psychological Science*, *7*(2), 54–60. https://doi.org/10.1111/1467-8721.ep13175642

Krauss, R. M., Chen, Y., & Gottesman, R. F. (2000). Lexical gestures and lexical access: a process model. In D. McNeill (Ed.), *Language and Gesture* (pp. 261–283). Cambridge University Press. https://doi.org/10.1017/cbo9780511620850.017

Krönke, K. M., Mueller, K., Friederici, A. D., & Obrig, H. (2013). Learning by doing? The effect of gestures on implicit retrieval of

newly acquired words. *Cortex*, *49*(9), 2553–2568. https://doi.org/10.1016/j.cortex.2012.11.016

Kushch, O. (2018). *Beat gestures and prosodic prominence: Impact on learning*. Universitat Pompeu Fabra (Spain).

Kushch, O., Igualada, A., & Prieto, P. (2018). Prominence in speech and gesture favour second language novel word learning. *Language, Cognition and Neuroscience*, *33*(8), 992–1004. https://doi.org/10.1080/23273798.2018.1435894

Lambacher, S. G., Martens, W. L., Kakehi, K., Marasinghe, C. A., & Molholt, G. (2005). The effects of identification training on the identification and production of American English vowels by native speakers of Japanese. *Applied Psycholinguistics*, *26*(2), 227–247. https://doi.org/10.1017.S0142716405050150

Law, L. N. C., & Zentner, M. (2012). Assessing Musical Abilities Objectively: Construction and Validation of the Profile of Music Perception Skills. *PLoS ONE*, *7*(12). https://doi.org/10.1371/journal.pone.0052508

Lee, B., Plonsky, L., & Saito, K. (2020). The effects of perception- vs. production-based pronunciation instruction. *System*, *88*, 102185. https://doi.org/10.1016/j.system.2019.102185

Lee, J., Jang, J., & Plonsky, L. (2015). The Effectiveness of Second Language Pronunciation Instruction: A Meta-Analysis. *Applied Linguistics*, *36*(3), 345–366. https://doi.org/10.1093/applin/amu040

Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2020). *Emmeans: Estimated marginal means, Aka Least-Squares means*. R package 1.5.1. https://cran.r-project.org/package=emmeans

Levis, J., & Pickering, L. (2004). Teaching intonation in discourse using speech visualization technology. *System*, *32*(4 SPEC.ISS.), 505–524. https://doi.org/10.1016/j.system.2004.09.009

Levy, E. S., & Law, F. F. (2010). Production of French vowels by American-English learners of French: Language experience, consonantal context, and the perception-production relationship. *The Journal of the Acoustical Society of America*, *128*(3), 1290–1305. https://doi.org/10.1121/1.3466879

Li, M., & Dekeyser, R. (2017). Perception practice, production practice, and musical ability in L2 Mandarin tone-word learning. *Studies in Second Language Acquisition*, *39*(4), 593–620. https://doi.org/10.1017/S0272263116000358

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

Li, P., Xi, X., Baills, F., & Prieto, P. (2020). Appropriately performing hand gestures cueing phonetic features facilitates simultaneous speech imitation in an L2. *Proceedings of the 7th Gesture and Speech in Interaction GESPIN 7*.

Li, P., Xi, X., Baills, F., & Prieto, P. (2021). Training non-native aspirated plosives with hand gestures : learners ' gesture performance matters. *Language, Cognition and Neuroscience*, 1–16. https://doi.org/10.1080/23273798.2021.1937663

Lightbown, P. M. (2008). Transfer appropriate processing as a model for classroom second language acquisition. In Z. Han (Ed.), *Understanding Second Language Process* (Vol. 27, p. 44). Multilingual Matters.

Lindblom, B. (1963). Spectrographic study of vowel reduction. *The Journal of the Acoustical Society of America*, *35*(11), 1773–1781.

Llanes-Coromina, J., Prieto, P., & Rohrer, P. L. (2018). Brief training with rhythmic beat gestures helps L2 pronunciation in a reading aloud task. In K. Klessa, J. Bachan, A. Wagner, M. Karpiński, & D. Śledziński (Eds.), *Proceedings of the International Conference on Speech Prosody* (pp. 498–502). https://doi.org/10.21437/SpeechProsody.2018-101

Llompart, M., & Reinisch, E. (2019). Imitation in a second language relies on phonological categories but does not reflect the productive usage of difficult sound contrasts. *Language and Speech*, *62*(3), 594–622. https://doi.org/10.1177/0023830918803978

Lockhart, R. S. (2002). Levels of processing, transfer-appropriate processing, and the concept of robust encoding. *Memory*, *10*(5–6), 397–403. https://doi.org/10.1080/09658210244000225

Lord, G. (2005). (How) Can we teach foreign language pronunciation? On the effects of a Spanish phonetics course. *Hispania*, *88*(3), 557. https://doi.org/10.2307/20063159

Lord, G. (2008). Podcasting communities and second language pronunciation. *Foreign Language Annals*, *41*(2), 364–379. https://doi.org/10.1111/j.1944-9720.2008.tb03297.x

Lüdecke, D., Makowski, D., Waggoner, P., & Patil, I. (2019). *Performance: Assessment of regression models performance*. R package 1.5.1.

Macedonia, M. (2019). Embodied learning: Why at school the mind needs the body. *Frontiers in Psychology*, *10*(October), 1–8. https://doi.org/10.3389/fpsyg.2019.02098

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*(2), 74–88. https://doi.org/10.1111/mbe.12047

Macedonia, M., & Knösche, T. R. (2011). Body in mind: How gestures empower foreign language learning. *Mind, Brain, and Education*, *5*(4), 196–211. https://doi.org/10.1111/j.1751-228X.2011.01129.x

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

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*(NOV), 1–11. https://doi.org/10.3389/fnsys.2013.00084

Martinie, B., & Wachs, S. (2006). *Phonétique en Dialogues*. CLE International SEJER.

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal About Thought*. University of Chicago Press.

Melnik-Leroy, G. A., Turnbull, R., & Peperkamp, S. (2021). On the relationship between perception and production of L2 sounds: Evidence from Anglophones' processing of the French /u/–/y/ contrast. *Second Language Research*. https://doi.org/10.1177/0267658320988061

Missaglia, F. (2007). Prosodic training for adult Italian learners of German: the Contrastive Prosody Method. In J. Trouvain & U. Gut (Eds.), *Non-Native Prosody: Phonetic Description and*

Teaching Practice (Issue section 4). De Gruyter Mouton. https://doi.org/10.1515/9783110198751.2.237

Missaglia, F. (1999). Contrastive prosody in SLA. An empirical study with adult Italian learners of German. In J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, & A. C. Bailey (Eds.), *Proceedings of the 14th International Congress of Phonetic Sciences* (pp. 551–554). https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS1999/papers/p14_0551.pdf

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

Mizera, G. J. (2006). *Working memory and L2 oral fluency*. University of Pittsburgh.

Mo, Y., Cole, J., & Hasegawa-Johnson, M. (2009). Prosodic effects on vowel production: Evidence from formant structure. In International Speech Communication Assosication (Ed.), *Proceedings of the 10th Annual Conference of the International Speech Communication Association 2009* (pp. 2535–2538).

Mora, J. C., & Levkina, M. (2018). Training vowel perception through map tasks: The role of linguistic and cognitive complexity. In J. Levis (Ed.), *Proceedings of the 9th Pronunciation in Second Language Learning and Teaching conference* (pp. 151–162). Iowa State University.

Morett, L. M. (2014). When Hands Speak Louder Than Words: The Role of Gesture in the Communication, Encoding, and Recall of Words in a Novel Second Language. *The Modern Language Journal*, *98*(3), 834–853. https://doi.org/10.1111/j.1540-4781.2014.12125.x

Morett, L. M. (2018). In hand and in mind: Effects of gesture production and viewing on second language word learning. *Applied Psycholinguistics*, *39*(2), 355–381. https://doi.org/10.1017/S0142716417000388

Morett, L. M., & Chang, L. Y. (2015). Emphasising sound and meaning: pitch gestures enhance Mandarin lexical tone acquisition. *Language, Cognition and Neuroscience*, *30*(3), 347–353. https://doi.org/10.1080/23273798.2014.923105

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*(5), 519–533. https://doi.org/10.1016/S0022-5371(77)80016-9

Munro, M. J., & Derwing, T. M. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech: The role of speaking rate. *Studies in Second Language Acquisition*, *23*(4), 451–468. https://doi.org/10.1017/s0272263101004016

Munro, M. J., & Derwing, T. M. (2006). The functional load principle in ESL pronunciation instruction: An exploratory study. *System*, *34*(4), 520–531. https://doi.org/10.1016/j.system.2006.09.004

Munro, M. J., & Derwing, T. M. (2015). A prospectus for pronunciation research in the 21st century. *Journal of Second Language Pronunciation*, *1*(1), 11–42. https://doi.org/10.1075/jslp.1.1.01mun

Myung, J. Y., Blumstein, S. E., & Sedivy, J. C. (2006). Playing on the typewriter, typing on the piano: Manipulation knowledge of objects. *Cognition*, *98*(3), 223–243. https://doi.org/10.1016/j.cognition.2004.11.010

Nagle, C. L. (2018). Examining the temporal structure of the perception–production link in second language acquisition: A longitudinal study. *Language Learning*, *68*(1), 234–270. https://doi.org/10.1111/lang.12275

O'Brien, I., Segalowitz, N., Collentine, J. O. E., & Freed, B. (2006). Phonological memory and lexical, narrative, and grammatical skills in second language oral production by adult learners. *Applied Psycholinguistics*, *27*(3), 377–402. https://doi.org/10.1017.S0142716406060322

Odisho, E. Y. (2007). A Multisensory , Multicognitive Approach to Teaching Pronunciation. *Linguística - Revista de Estudos Linguísticos Da Universidade Do Porto - Vol. 2 - 2007, Pp. 3-28*, *2*, 3–28.

Odisho, E. Y. (2014). *Pronunciation is in the Brain, not in the Mouth:A Cognitive Approach to Teaching it*. Gorgias Press LLC.

Olson, D. J. (2014). Benefits of visual feedback on segmental production in the L2 classroom. *Language, Learning and Technology*, *18*(3), 173–192. https://doi.org/http://dx.doi.org/10125/44389

Ozakin, A. S., Xi, X., Li, P., & Prieto, P. (2021). Thanks or tanks: Training with tactile cues facilitates the pronunciation of non-native English interdental consonants. *Language Learning and Development*.

Paas, F., & Sweller, J. (2012). An evolutionary upgrade of Cognitive Load Theory: Using the human motor system and collaboration to support the learning of complex cognitive tasks. *Educational Psychology Review*, *24*(1), 27–45. https://doi.org/10.1007/s10648-011-9179-2

Padgett, J., & Tabain, M. (2005). Adaptive dispersion theory and phonological vowel reduction in Russian. *Phonetica*, *62*(1), 14–54.

Paivio, A. (1991). Dual Coding Theory: Retrospect and current status. *Canadian Journal of Psychology*, *45*(3), 255–287. http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,url,cookie,uid&an=1992-07881-001&db=psyh&scope=site&site=ehost

Paula Roncaglia-Denissen, M., Roor, D. A., Chen, A., & Sadakata, M. (2016). The enhanced musical rhythmic perception in second language learners. *Frontiers in Human Neuroscience*, *10*(June), 1–10. https://doi.org/10.3389/fnhum.2016.00288

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Peterson, R. A., & Cavanaugh, J. E. (2019). Ordered quantile normalization: a semiparametric transformation built for the cross-validation era. *Journal of Applied Statistics*, 1–16. https://doi.org/10.1080/02664763.2019.1630372

Ping, R., & Goldin-Meadow, S. (2010). Gesturing saves cognitive resources when talking about nonpresent objects. *Cognitive Science*, *34*(4), 602–619. https://doi.org/10.1111/j.1551-6709.2010.01102.x

Posedel, J., Emery, L., Souza, B., & Fountain, C. (2012). Pitch perception, working memory, and second-language phonological production. *Psychology of Music*, *40*(4), 508–517. https://doi.org/10.1177/0305735611415145

Post, L. S., van Gog, T., Paas, F., & Zwaan, R. A. (2013). Effects of simultaneously observing and making gestures while studying grammar animations on cognitive load and learning. *Computers in Human Behavior*, *29*(4), 1450–1455. https://doi.org/https://doi.org/10.1016/j.chb.2013.01.005

Prieto, P., Borràs-Comes, J., Cabré, T., Crespo-Sendra, V., Mascaró, I., Roseano, P., Sichel-Bazin, R., & del Mar Vanrell, M. (2015). Intonational phonology of Catalan and its dialectal varieties. In S. Frota & P. Prieto (Eds.), *Intonation in Romance* (pp. 9–62). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199685332.003.0002

Prieto, P., Cravotta, A., Kushch, 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.), *9th International Conference on Speech Prosody 2018* (pp. 201–205).

Pulvermüller, F., Hauk, O., Nikulin, V. V., & Ilmoniemi, R. J. (2005). Functional links between motor and language systems. *European Journal of Neuroscience*, *21*(3), 793–797. https://doi.org/10.1111/j.1460-9568.2005.03900.x

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

Racine, I., & Detey, S. (2019). Production of French close rounded vowels by Spanish learners. In M. Gibson & J. Gil (Eds.), *Romance Phonetics and Phonology* (pp. 381–394). Oxford University Press. https://doi.org/10.1093/oso/9780198739401.003.0019

Ramírez Verdugo, D. (2006). A study of intonation awareness and learning in non-native speakers of english. *Language Awareness*, *15*(3), 141–159. https://doi.org/10.2167/la404.0

Reiterer, S. M. (2019). Neuro-psycho-cognitive markers for pronunciation/speech imitation as language aptitude. In Z. (Edward) Wen, P. Skehan, A. Biedroń, S. Li, & R. L. Sparks (Eds.), *Language Aptitude* (pp. 277–298). Routledge.

Renard, R. (1979). *Introduction à la Méthode Verbo-tonale de Correction Phonétique* (3rd ed.). Didier Érudition - Centre International de Phonétique Appliquée a Mons.

Renard, R. (2002). Une phonétique immergée. In R. Renard (Ed.), *Apprentissage d'une Langue Étrangère/Seconde Vol.2 La Phonétique Verbo-tonale* (pp. 12–24). De Boeck Université.

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, *20*(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002

Rizvanović, N. (2018). Motivation and Personality in Language Aptitude. In S. M. Reiterer (Ed.), *Exploring Language Aptitude: Views from Psychology, the Language Sciences, and Cognitive Neuroscience* (pp. 101–116). Springer.

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

Roberge, C., Kimura, M., & Kawaguchi, Y. (1996). *Nihongo no hatsuon shidoo: VThoo no riron to jissai [Pronunciation training for Japanese: Theory and practice of the VT method]*. Bonjinsha.

Rognoni, L., & Busà, M. G. (2014). Testing the effects of segmental and suprasegmental pphonetic cues in foreign accent rating : An experiment using prosody transplantation. *Proceedings of the International Symposium on the Acquisition of Second Language Speech Concordia Working Papers in Applied Linguistics*, *5*, 547–560. http://doe.concordia.ca/copal/documents/35_Rognoni_Busa_Vol5.pdf

Rosborough, A. A. (2010). *Gesture as an act of meaning-making: An eco-social perspective of a sheltered-English second grade classroom* [University of Nevada, Las Vegas]. http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=ED525768&site=ehost-live%5Cnhttp://gateway.proquest.com/openurl?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:dissertation&res_dat=xri:pqdiss&rft_dat=xri:pqdiss:3440006

Rota, G., & Reiterer, S. M. (2009). Cognitive aspects of pronunciation talent. In G. Dogil & S. M. Reiterer (Eds.), *Language Talent and Brain Activity* (pp. 67–96). Walter de Gruyter. https://doi.org/10.1515/9783110215496.67

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

Rusiewicz, H. L., Shaiman, S., Iverson, J. M., & Szuminsky, N. (2013). Effects of prosody and position on the timing of deictic gestures. *Journal of Speech, Language, and Hearing Research*, *56*(2), 458–470. https://doi.org/10.1044/1092-4388(2012/11-0283)

Rusiewicz, H. L., Shaiman, S., Iverson, J. M., & Szuminsky, N. (2014). Effects of perturbation and prosody on the coordination of speech and gesture. *Speech Communication*, *57*, 283–300. https://doi.org/10.1016/j.specom.2013.06.004

Sabkay71. (2011). *BBC's Proms Hedwig's Theme from Harry Potter*. YouTube. https://www.youtube.com/watch?v=GTXBLyp7_Dw

Sadakata, M., & Sekiyama, K. (2011). Enhanced perception of various linguistic features by musicians: A cross-linguistic study. *Acta Psychologica*, *138*(1), 1–10. https://doi.org/10.1016/j.actpsy.2011.03.007

Saito, K., & Lyster, R. (2012). Effects of form-focused instruction and corrective feedback on L2 pronunciation development of /ɹ/ by Japanese learners of English. *Language Learning*, *62*(2), 595–633. https://doi.org/10.1111/j.1467-9922.2011.00639.x

Saito, K., & Munro, M. J. (2014). The early phase of /ɹ/ production development in adult Japanese learners of English. *Language and Speech*, *57*(4), 451–469. https://doi.org/10.1177/0023830913513206

Saito, K., & Plonsky, L. (2019). Effects of Second Language Pronunciation Teaching Revisited: A Proposed Measurement Framework and Meta-Analysis. *Language Learning*, *69*(3), 652–708. https://doi.org/10.1111/lang.12345

Saito, K., Trofimovich, P., & Isaacs, T. (2016). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics*, *37*(2), 217–240. https://doi.org/10.1017/S0142716414000502

Saito, K., Trofimovich, P., & Isaacs, T. (2017). Using listener judgments to investigate linguistic influences on L2 comprehensibility and accentedness: A validation and generalization Study. *Applied Linguistics*, *38*(4), 439–462. https://doi.org/10.1093/applin/amv047

Saito, Y., & Saito, K. (2017). Differential effects of instruction on the development of second language comprehensibility, word stress, rhythm, and intonation: The case of inexperienced Japanese EFL learners. *Language Teaching Research*, *21*(5), 589–608. https://doi.org/10.1177/1362168816643111

Santiago, F. (2021). L'accentuation contribue-t-elle à l'acquisition du contraste arrondi vs non-arrondi des voyelles orales en français langue étrangère ? *Etudes de Linguistique Appliquée*, 74–90.

Santiago, F., & Mairano, P. (2019). Prosodic effects on L2 French vowels : a corpus-based investigation. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences* (pp. 1084–1088). Australasian Speech Science and Technology Association Inc.

Schwab, S., & Goldman, J. P. (2018). MIAPARLE: Online training for discrimination and production of stress contrasts. In K. Klessa, J. Bachan, A. Wagner, M. Karpiński, & D. Śledziński (Eds.), *Proceedings of the International Conference on Speech Prosody* (pp. 572–576). https://doi.org/10.21437/SpeechProsody.2018-116

Seferoğlu, G. (2005). Improving students' pronunciation through accent reduction software. *British Journal of Educational Technology*, *36*(2), 303–316. https://doi.org/10.1111/j.1467-8535.2005.00459.x

Segalowitz, N. (1997). Individual differences in second language acquisition. In *Tutorials in Bilingualism: Psycholinguistic*

Perspectives. (pp. 85–112). Lawrence Erlbaum Associates Publishers.

Segalowitz, N. (2000). Automaticity and attentional skill in fluent performance. In H. Riggenbach (Ed.), *Perspectives on Fluency* (pp. 200–219). University of Michigan Press.

Shapiro, L. (2011). *Embodied Cognition*. Routledge. https://doi.org/10.4337/9781848447424.00008

Shapiro, L. (2014). The routledge handbook of embodied cognition. In *The Routledge Handbook of Embodied Cognition*. https://doi.org/10.4324/9781315775845

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

Skulmowski, A., & Rey, G. D. (2017). Measuring cognitive load in embodied learning settings. *Frontiers in Psychology*, *8*(AUG), 1–6. https://doi.org/10.3389/fpsyg.2017.01191

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

So, W. C., Sim Chen-Hui, C., & Low Wei-Shan, J. (2012). Mnemonic effect of iconic gesture and beat gesture in adults and children: Is meaning in gesture important for memory recall? *Language and Cognitive Processes*, *27*(5), 665–681. https://doi.org/10.1080/01690965.2011.573220

Suzukida, Y., & Saito, K. (2019). Which segmental features matter for successful L2 comprehensibility? Revisiting and generalizing the

pedagogical value of the functional load principle. *Language Teaching Research*. https://doi.org/10.1177/1362168819858246

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285. https://doi.org/10.1016/0364-0213(88)90023-7

Sweller, J., Merrienboer, J. J. G. van, & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, *10*(3), 251–296. https://doi.org/10.1023/A

Sweller, N., Shinooka-Phelan, A., & Austin, E. (2020). The effects of observing and producing gestures on Japanese word learning. *Acta Psychologica*, *207*(April), 103079. https://doi.org/10.1016/j.actpsy.2020.103079

Tajima, K., Kato, H., Rothwell, A., Akahane-Yamada, R., & Munhall, K. G. (2008). Training English listeners to perceive phonemic length contrasts in Japanese. *The Journal of the Acoustical Society of America*, *123*(1), 397–413. https://doi.org/10.1121/1.2804942

Taleghani-Nikazm, C. (2008). Gestures in foreign language classrooms: An empirical analysis of their organization and function. In M. Boweles, R. Foote, S. Perpiñán, & R. Bhatt (Eds.), *Selected Proceedings of the 2007 Second Language Research Forum* (Issue July, pp. 229–238). Cascadilla Proceedings Project. http://www.lingref.com/cpp/slrf/2007/paper1747.pdf

Tanner, M. W., & Landon, M. M. (2009). The effects of computer-assisted pronunciation readings on ESL learners' use of pausing, stress, intonation, and overall comprehensibility. *Language Learning & Technology*, *13*(3), 51–65. http://llt.msu.edu/vol13num3/tannerlandon.pdf

Tellier, M. (2008). The effect of gestures on second language memorisation by young children. *Gesture*, *8*(2), 219–235. https://doi.org/10.1075/bct.28.06tel

Traunmüller, H. (1990). Analytical expressions for the tonotopic sensory scale. *Journal of the Acoustical Society of America*, *88*(1), 97–100. https://doi.org/10.1121/1.399849

Trofimovich, P., & Baker, W. (2006). Learning second language suprasegmentals: Effects of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition*, *28*(1), 1–30. http://eprints.whiterose.ac.uk/72874/

Trofimovich, P., & Isaacs, T. (2012). Disentangling accent from comprehensibility. *Bilingualism*, *15*(4), 905–916. https://doi.org/10.1017/S1366728912000168

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*(1), 259–280. https://doi.org/10.1111/j.1467-9922.2010.00611.x

van Gog, T., Paas, F., Marcus, N., Ayres, P., & Sweller, J. (2009). The mirror neuron system and observational learning: Implications for the effectiveness of dynamic visualizations. *Educational Psychology Review*, *21*(1), 21–30. https://doi.org/10.1007/s10648-008-9094-3

van Maastricht, L., Hoetjes, M., & van Drie, E. (2019). Do gestures during training facilitate L2 lexical stress acquisition by Dutch learners of Spanish? In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences* (pp. 6–10). Australasian Speech Science and Technology Association Inc. https://doi.org/10.21437/avsp.2019-2

van Merriënboer, J. J. G., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review*, *17*(2), 147–177. https://doi.org/10.1007/s10648-005-3951-0

Walker, R. (2011). *Vowel Patterns in Language*. Cambridge university Press.

Wang, H., Mok, P., & Meng, H. (2016). Capitalizing on musical rhythm for prosodic training in computer-aided language learning. *Computer Speech and Language*, *37*, 67–81. https://doi.org/10.1016/j.csl.2015.10.002

Wang, X. (2020). Segmental versus suprasegmental: Which one is more important to teach? *RELC Journal*, 397–412. https://doi.org/10.1177/0033688220925926

Wellsby, M., & Pexman, P. M. (2014). Developing embodied cognition: Insights from children's concepts and language processing. *Frontiers in Psychology*, *5*(MAY), 1–10. https://doi.org/10.3389/fpsyg.2014.00506

Weltens, B., & de Bot, K. (1984a). The visualisation of pitch contours: Some aspects of its effectiveness in teaching foreign intonation. *Speech Communication*, *3*(2), 157–163. https://doi.org/10.1016/0167-6393(84)90037-2

Weltens, B., & de Bot, K. (1984b). Visual feedback of intonation II: Feedback delay and quality of feedback. *Language and Speech*, *27*(1), 79–88.

Wen, Z. (2018). *Working Memory and Second Language Learning: Towards an Integrated Approach*. Multilingual Matters.

Wheeler, M. W. (2005). *The Phonology of Catalan*. Oxford University Press.

Whitfield, J. A., Reif, A., & Goberman, A. M. (2018). Voicing contrast of stop consonant production in the speech of individuals with Parkinson disease ON and OFF dopaminergic medication. *Clinical Linguistics and Phonetics*, *32*(7), 587–594. https://doi.org/10.1080/02699206.2017.1387816

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin and Review*, *9*(4), 625–636. https://doi.org/10.3758/BF03196322

Wlomainck, P. (2002). Le travail du rythme et de l'intonation dans l'apprentissage d'une langue étrangère. In R. Renard (Ed.), *Apprentissage d'une Langue Étrangère/Seconde Vol.2 La Phonétique Verbo-tonale* (pp. 156–161). De Boeck Université.

Wong, P. C. M., & Perrachione, T. K. (2007). Learning pitch patterns in lexical identification by native English-speaking adults. *Applied Psycholinguistics*, *28*(4), 565–585. https://doi.org/10.1017/S0142716407070312

Woods, D. L., Kishiyama, M. M., Yund, E. W., Herron, T. J., Edwards, B., Poliva, O., Hink, R. F., & Reed, B. (2011). Improving digit span assessment of short-term verbal memory. *Journal of Clinical and Experimental Neuropsychology*, *33*(1), 101–111. https://doi.org/10.1080/13803395.2010.493149

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

Yang, J., & Shu, H. (2016). Involvement of the motor system in comprehension of non-literal action language: A meta-analysis study. *Brain Topography*, *29*(1), 94–107. https://doi.org/10.1007/s10548-015-0427-5

Yang, Y. (2016). *Improving the English speaking kills and phonological working memory of Chinese primary EFL learners with a verbotonal-based approach* [Suranaree University of Technology]. https://doi.org/10.13140/RG.2.2.23991.83362

Yeo, A., Ledesma, I., Nathan, M. J., Alibali, M. W., & Church, R. B. (2017). Teachers'gestures and students'learning：sometimes"hands off"is better. *Cognitive Research: Principles and Implications*, *2*(1). https://doi.org/10.1186/s41235-017-0077-0

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

Zampini, M. L. (1998). The relationship between the production and perception of L2 Spanish stops. *Texas Papers in Foreign Language Education*, *3*(3), 85–100.

Zhang, F. Z. (2006). *The Teaching of Mandarin Prosody: A Somatically-Enhanced Approach For Second Language Learners* [University of Canberra, Australia]. https://doi.org/10.6084/M9.FIGSHARE.1189254

Zhang, R., & Yuan, Z. M. (2020). Examining the effects of explicit pronunciation instruction on the development of L2 pronunciation. *Studies in Second Language Acquisition*, *42*(4), 905–918. https://doi.org/10.1017/S0272263120000121

Zhang, Y. (2002). Shoushi zai yuyin jiaoxue zhong de zuoyong [The importance of using gestures in pronunciaiton teaching]. *Language Teaching and Linguistic Studies*, *6*, 51–56.

Zhang, Y., Baills, F., & Prieto, P. (2020). Hand-clapping to the rhythm of newly learned words improves L2 pronunciation: Evidence from training Chinese adolescents with French words. *Language Teaching Research*, *24*(5), 666–689. https://doi.org/10.1177/1362168818806531

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

Zheng, A., Hirata, Y., & Kelly, S. D. (2018). Exploring the effects of imitating hand gestures and head nods on L1 and L2 Mandarin tone production. *Journal of Speech, Language, and Hearing Research*, *61*(9), 2179–2195. https://doi.org/10.1044/2018_JSLHR-S-17-0481

Zielinski, B. (2015). The segmental/suprasegmental debate. In M. Reed & J. M. Levis (Eds.), *The Handbook of English Pronunciation* (pp. 397–412). John Wiley & Sons.

APPENDIX A: CHAPTER 2

Table 1

Catalan 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
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

Table 2

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 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

APPENDIX B: CHAPTER 3

Table 1

Stimuli for the Pronunciation Training Session

Consonant pairs	Training word pairs	Catalan transcription [a]	English gloss
/p/ - /pʰ/	*bíyán/píyán*	*pi yan/pʰi yan*	rhinitis/dermatitis
	bōfù/pōfù	*puo fu/pʰuo fu*	disbursement/shrew
/t/ - /tʰ/	*dānliàn/tānliàn*	*tan lian/tʰan lian*	one-sided love/greedy
	dúlì/túlì	*tu li/tʰu li*	independence/figure
/k/ - /kʰ/	*guìyang/kuìyáng*	*kuei yang/kʰuei yang*	name of a city/ulcer
	gǒuliáng/kǒuliáng	*kou liang/kʰou liang*	dog food/ration

a To facilitate reading by Catalan speakers without any knowledge of pinyin, the words were transcribed in accordance with Catalan orthography. All the tonal markers were left out in the Catalan transcription, given that lexical tones were not of interest in this study.

Table 2

Stimuli for the Identification Task

Consonant pairs	Word pairs	Catalan transcription [a]	English Gloss
Trained			
/p/ - /pʰ/	*bíyán/píyán*	*pi yan/pʰi yan*	rhinitis/dermatitis
/t/ - /tʰ/	*dānliàn/tānliàn*	*tan lian/tʰan lian*	one-side love/greedy
/k/ - /kʰ/	*guìyang/kuìyáng*	*kuei yang/kʰuei yang*	name of a city/ulcer
Untrained			
/p/ - /pʰ/	*báiliàn/páiliàn*	*pai lian/pʰai lian*	white silk/rehearsal
/t/ - /tʰ/	*dōngfēng/tōngfēng*	*tong feng/tʰong feng*	east wind/ventilation
/k/ - /kʰ/	*gǔlì/kǔlì*	*ku li/kʰu li*	encouragement/labour

a This transcription was for Catalan speakers to choose the correct answer, which was consistent with the transcription used in the pronunciation training session.

Table 3

Stimuli for the Imitation Task

Consonant pairs	Word pairs	English Gloss
Trained		
/p/ - /pʰ/	*bōfù/pōfù*	disbursement/shrew
/t/ - /tʰ/	*dúlì/túlì*	independence/figure
/k/ - /kʰ/	*gǒuliáng/kǒuliáng*	dog food/ration
Untrained		
/p/ - /pʰ/	*báshǒu/páshǒu*	handle/thief
/t/ - /tʰ/	*dàolù/tàolù*	road/strategy
/k/ - /kʰ/	*gōnglíng/kōnglíng*	seniority/ethereal

Table 4

Stimuli for the Vocabulary Training Session

Consonant pairs	Word pairs	Catalan translation	English gloss
/p/ - /pʰ/	*bí/pí*	nas/pell	nose/skin
	bái/pái	blanc/fila	white/row
/t/ - /tʰ/	*dù/tù*	ventre/conill	stomach/rabbit
	dàn/tàn	ou/carbó	egg/carbon
/k/ - /kʰ/	*guāng/kuāng*	llum/cistella	light/basket
	guī/kuī	tortuga/casc	turtle/helmet

Table 5

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 Chinese?
Do you want to learn Chinese? Evaluate your motivation from 1 (not at all) to 9 (very much)

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?
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
How often do you 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

Table 6

Summary of the Mixed-Effects Models Involving Familiarity (Trained vs. Untrained) for the Identification Score, VOT Ratio, and Pronunciation Score

	Identification			VOT ratio			Pronunciation		
	χ^2	*df*	*p*	χ^2	*df*	*p*	χ^2	*df*	*p*
Aspiration									
	1.06	1	.304	43.37	1	<.001	0.01	1	.927
Gesture Performance									
	0.12	3	.989	5.24	3	.155	1.69	3	.639
Test									
	5.21	2	.074	8.16	2	.017	43.32	2	<.001
Familiarity									

0.06	**1**	**.813**	**3.42**	**1**	**.064**	**0.56**	**1**	**.454**
Aspiration × Gesture performance								
6.03	3	.110	2.99	3	.393	1.94	3	.585
Aspiration × Test								
1.38	2	.501	22.31	2	<.001	25.22	2	<.001
Gesture performance × Test								
7.93	6	.244	14.81	6	.022	18.18	6	.006
Aspiration × Familiarity								
0.12	1	.732	0.01	1	.911	0.01	1	.903
Gesture performance × Familiarity								
4.77	3	.190	5.91	3	.116	7.07	3	.070
Test × Familiarity								
1.83	2	.401	0.63	2	.731	1.37	2	.503
Aspiration × Gesture performance × Test								
2.40	6	.880	19.67	6	.003	7.95	6	.242
Aspiration × Gesture performance × Familiarity								
0.88	3	.830	8.28	3	.040	1.17	3	.759
Aspiration × Test × Familiarity								
3.46	2	.178	5.45	2	.066	2.01	2	.366
Gesture performance × Test × Familiarity								
8.48	**6**	**.205**	**4.22**	**6**	**.647**	**4.41**	**6**	**.621**
Aspiration × Gesture performance × Test × Familiarity								
5.88	**6**	**.437**	**3.23**	**6**	**.780**	**2.40**	**6**	**.880**

Note. The main effect of familiarity (untrained items vs. trained items) and relevant interactions involving familiarity are in boldface.

Table 7

Means, Standard Deviations, and the Statistical Results of Linear Model Analysis in Individual Differences with Gesture Performance as the Fixed Factor in the Pronunciation Training Session

NG	PPG	WPG	C	*F*(3)	*p*

	M (SD)	*M (SD)*	*M (SD)*	*M (SD)*		
Age	19.14 (1.53)	18.67 (1.18)	20.00 (2.00)	19.89 (1.76)	2.18	.100
DS	6.36 (1.13)	6.35 (0.93)	6.86 (1.43)	6.14 (0.86)	0.76	.520
MES	10.03 (4.49)	9.47 (3.23)	9.64 (4.92)	10.89 (6.62)	0.05	.986
MS	6.72 (1.75)	5.87 (2.00)	5.93 (1.90)	6.11 (1.05)	1.10	.354

Note. DS = Digit span score; MES = Musical experience score; MS = Motivation score; NG = No Gesture; PPG = Poorly Performed Gesture; WPG = Well Performed Gesture

Table 8

Means, Standard Deviations, and the Statistical Results of Linear Model Analysis in Individual Differences with Gesture Performance as the Fixed Factor in the Vocabulary Training Session

	NG	PPG	WPG	*F*(2)	*p*
	M (SD)	*M (SD)*	*M (SD)*		
Age	19.14 (1.53)	18.88 (1.45)	19.92 (1.98)	1.54	.223
DS	6.36 (1.13)	6.46 (0.90)	6.80 (1.56)	0.51	.601
MES	10.03 (4.49)	9.06 (3.19)	10.25 (5.12)	0.19	.830
MS	6.72 (1.75)	6.18 (1.78)	5.50 (2.11)	1.86	.165

Note. DS = Digit span score; MES = Musical experience score; MS = Motivation score; NG = No Gesture; PPG = Poorly Performed Gesture; WPG = Well Performed Gesture

APPENDIX C: CHAPTER 4

Table 1

Dialogues Used in the Training Sessions and for the Dialogue-Reading Task

Session	Dialogue

Session 1

Tu ne vas pas t'amuser avec cet hurluberlu !

- Eh ! Sal**u**t L**u**cie !
- Sal**u**t M**u**riel. Mais ch**u**t ! Ne parle pas si fort. T**u** es folle ! Où vas-t**u** ?
- Faire un tour dans la r**u**e. Et toi, pourquoi es-t**u** descend**u**e ?
- Parce que j'aime la r**u**e, c'est tout.
- C'est tout ? T**u** es s**û**re ?
- Oui, je suis s**û**re. Pourquoi ?
- Je te trouve pl**u**tôt triste. Allez, viens ! On sort **u**ne min**u**te.
- T**u** as v**u** tous ces nuages ? Il va pl**eu**voir, c'est s**û**r. On n'a pas **eu** de pluie depuis le déb**u**t d**u** mois de mars...
- Pourquoi t**u** es si triste, L**u**cie ? T**u** n'aimes pas ces vacances ?
- M**u**riel, c'est trop d**u**r. L**u**c ne me parle pl**u**s depuis sept jours...
- T**u** sais, c'est un vrai sauvage, ce L**u**c. T**u** ne vas pas t'am**u**ser avec cet h**u**rl**u**berl**u** !

(Adapted from Martinie & Wachs, 2006, p. 15)

Session 2

Tu as eu peur ?

- Je ne trouve pas la ceint**u**re de séc**u**rité.
- Normal ! Cette voit**u**re est de quatre-vingt-d**eu**x, il n'y avait pas de ceint**u**re à l'époque.
- Ah bon ? Mais c'est très danger**eu**x !
- Danger**eu**x ? Pas d**u** tout !
- Mon **œ**il… À quelle h**eu**re est le rendez-vous chez le coiff**eu**r ?
- À 13 h**eu**res. Mais ne t'inquiète pas, c'est j**u**ste à côté. Profite des paysages !
- Oui, c'est très beau, mais ... je suis un p**eu** malade en voit**u**re.
- T**u** n'as pas à avoir p**eu**r : je suis le meill**eu**r chauff**eu**r de toute la M**eu**se !
- Je vois le vide, quelle horr**eu**r… Au secours !
- On est arrivés. T**u** as **eu** p**eu**r ?
- Non, non. J**u**ste un p**eu** mal au c**œu**r... Merci **Eu**génie, heim… Mais, je ne v**eu**x pas te retenir : je rentrerai s**eu**le…

(Adapted from Martinie & Wachs, 2006, p. 22)

Session 3

Une gentille vagabonde

- Bonjour, génér**eu**se demoiselle ! Vous avez bien **u**ne petite pièce pour **u**ne pauvre misér**eu**se !

- **Eu**h voyons voir... Ah, mais quel malh**eu**r ! Je ne trouve pl**u**s mon portef**eu**ille. Regardez : la poche de ma veste est vide !
- La chance n'est pas en ma fav**eu**r !
- Ah, je suis bien anxi**eu**se maintenant : j'ai un train de banli**eu**e à sept h**eu**res, et je n'ai pas de billet ! Comment vais-je trouver d**eu**x **eu**ros ?
- Un mouchoir, d**u** tabac, un vi**eu**x bout de ficelle ... Ah, voilà vos d**eu**x **eu**ros !
- Non, je ne v**eu**x pas ab**u**ser de votre amabilité.
- Si, si ! Prenez ! C'est de bon c**œu**r.
- Alors j'accepte bien volontiers. Merci beaucoup. Au revoir !

(Adapted from Martinie & Wachs, 2006, p. 44)

Untrained dialogue

Je travaille, moi !

- Je ne p**eu**x pas passer ! Soyez gentil, monsi**eu**r, dégagez le passage, par pitié !
- Dégager le passage ? S**û**rement pas. Je travaille, moi ! D'aill**eu**rs, c'est vous qui gênez. À vous de bouger votre voit**u**re.
- Je n'en crois ni mes y**eu**x, ni mes oreilles ! C'est l'h**eu**re de déj**eu**ner, et je suis déjà en retard... Que malh**eu**r !
- C'est bien dommage. Bon, allez, entend**u** ! Je suis prêt à vous arranger. Mais vous m'aidez à décharger ces cartons !
- Oh ! Espèce de m**u**fle ! Que vous êtes mal élevé !
- Des inj**u**res, à présent ? J**u**sque-là, je suis resté gentil. Mais pl**u**s question que je parte. Voyez-vous, je p**eu**x patienter toute la journée avec mon journal !

(Adapted from Martinie & Wachs, 2006, p.48)

Note. The target phonemes are in boldface.

Table 2

Sentences Included the Sentence Imitation Task

	Sentences
Trained	
/y/	
	Faire un tour dans la rue.
	Pourquoi es-tu descendue ?
	C'est tout ? Tu es sûre ?
	Je te trouve plutôt triste.
/ø/	
	Ah bon ? Mais, c'est très dangereux.
	Je ne veux pas te retenir.
	Comment vais-je trouver deux euros ?

	Je suis bien anxieuse maintenant.
/œ/	
	Je rentrai seule.
	Je suis le meilleur chauffeur.
	Si, si, prenez. C'est de bon cœur.
	La chance n'est pas en ma faveur.
Untrained	
/y/	À vous de bouger votre voiture.
/ø/	Je n'en crois, ni mes yeux, ni mes oreilles.
/œ/	D'ailleurs, c'est vous qui gênez.

Note. The target phonemes are in boldface.

Table 3

Pairwise Comparisons of F1 (Bark) of the Target Front Rounded Vowels by Test in the Dialogue-Reading Task

	Estimate	*SE*	*df*	*t* ratio	*p* value	Cohen's *d*
Pretest						
/y/-/ø/	-0.64	0.16	1001	-4.04	<.001	-0.97
/y/-/œ/	-1.53	0.16	1001	-9.71	<.001	-2.32
/ø/-/œ/	-0.90	0.16	1001	-5.46	<.001	-1.36
Posttest						
/y/-/ø/	-0.63	0.17	1001	-3.72	.001	-0.99
/y/-/œ/	-1.61	0.17	1001	-9.58	<.001	-2.54
/ø/-/œ/	-0.99	0.18	1001	-5.65	<.001	-1.56
Delayed posttest						
/y/-/ø/	-0.58	0.17	1001	-3.54	.001	-0.86
/y/-/œ/	-1.61	0.17	1001	-9.78	<.001	-2.38
/ø/-/œ/	-1.03	0.17	1001	-6.01	<.001	-1.51

Table 4

Pairwise Comparisons of F1 (Bark) of the Target Front Rounded Vowels by Test in the Sentence Imitation Task

	Estimate	*SE*	*df*	*t* ratio	*p* value	Cohen's *d*
Pretest						
/y/-/ø/	-0.69	0.22	1001	-3.10	.006	-1.02
/y/-/œ/	-1.69	0.22	1001	-7.65	<.001	-2.52
/ø/-/œ/	-1.01	0.23	1001	-4.38	<.001	-1.50
Posttest						
/y/-/ø/	-0.73	0.23	1001	-3.22	.004	-1.07
/y/-/œ/	-1.74	0.23	1001	-7.71	<.001	-2.55
/ø/-/œ/	-1.01	0.23	1001	-4.33	.001	-1.49
Delayed posttest						
/y/-/ø/	-0.64	0.21	1001	-3.05	.007	-1.07
/y/-/œ/	-1.73	0.21	1001	-8.18	<.001	-2.87
/ø/-/œ/	-1.09	0.22	1001	-4.95	<.001	-1.80

APPENDIX D: PUBLICATION LIST

The following publications are associated with this dissertation

Publications in peer-reviewed journals (JCR)

- Li, P., Xi, X., Baills, F. & Prieto, P. (2021, in press). Training non-native aspirated plosives with hand gestures: Learners' gesture performance matters. *Language Cognition and Neuroscience*. 10.1080/23273798.2021.1937663

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

- Xi, X., Li, P., Baills, F. & Prieto, P. (2020). Hand gestures facilitate the acquisition of novel phonemic contrasts when 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.

Publications in peer-reviewed conference proceedings

- Li, P., Xi, X., & Baills, F., Prieto, P. (2020). Appropriately performing hand gestures cueing phonetic features facilitates simultaneous speech

imitation in an L2. In *Proceedings of the 7th GeSpIn Conference,* GeSpIN 2020, 7-9 September 2020. KTH Royal Institute of Technology, Sweden.

- Xi, X., Li, P., Baills, F., & Prieto, P. (2020). Training the pronunciation of L2 novel phonetic features: A comparison of observing versus producing hand gestures. In *Proceedings of the 7th GeSpIn Conference,* GeSpIN 2020, 7-9 September 2020. KTH Royal Institute of Technology, Sweden.

Manuscripts under review or in preparation

- Li, P., Baills, F., Baqué, L., & Prieto, P. (under review). Embodied prosodic training helps improve not only accentedness but also vowel accuracy. *Language Teaching Research*.

- Ozakin, A., Xi, X., Li, P., & Prieto, P. (under review). Thanks or tanks: Training with tactile cues facilitates the pronunciation of non-native English interdental consonants. *Language, Learning and Development*.

- Xi. X., Li, P., Baills, F., & Prieto, P. (in prep.). Appropriate gesture performance helps the learning of novel segmental features more than observing gestures.