Learning novel phonemic contrasts in the L2: the effects of audiovisual phonetic training with visuospatial hand gestures

Xiaotong Xi

Supervisor: Dr. Pilar Prieto
#TABLE OF CONTENTS

Abstract ........................................................................................................... i
Resum .............................................................................................................. iii
Resumen ......................................................................................................... v

1. Introduction ................................................................................................. 1

1.1 Object of analysis ...................................................................................... 1

1.2 Literature Review ...................................................................................... 2

1.2.1 L2 pronunciation learning and teaching ............................................. 2

1.2.2 High Variability Phonetic Training (HVPT) ....................................... 4

1.2.3 Audiovisual phonetic training ............................................................. 6

1.2.4 Effects of hand gestures on language learning .................................. 8

1.2.4.1 The benefits of hand gesture for L2 acquisition ......................... 8

1.2.4.2 The benefits of visuospatial hand gesture for L2 prosodic learning . 10

1.2.4.3 The benefits of visuospatial hand gesture for L2 phonetic learning ... 13

2. Goals and Hypotheses .............................................................................. 21

2.1 Goals of the dissertation .......................................................................... 21

2.2 Hypotheses ............................................................................................... 22

2.2.1 Study 1 ............................................................................................... 22

2.2.2 Study 2 ............................................................................................... 23

2.2.2 Study 3 ............................................................................................... 24

3. Experimental studies ................................................................................ 25

3.1 Study 1: Training with a fist-to-open hand gesture on the learning of novel
aspiration contrasts ....................................................................................... 25

3.1.1 Research questions ............................................................................. 25

3.1.2 Participants .......................................................................................... 25

3.1.3 Materials ............................................................................................ 25

3.1.4 Experimental procedure ..................................................................... 28

3.1.5 Coding of data .................................................................................... 29

3.1.6 Statistical analyses ............................................................................ 31

3.1.7 Results ................................................................................................ 32

3.2 Study 2: Training with a palm-down finger-lift gesture for the learning of alveolar-
retroflex contrasts .................................................................36

3.2.1 Research questions ..........................................................36
3.2.2 Participants .................................................................37
3.2.3 Materials .................................................................37
3.2.4 Experimental procedure ...................................................37
3.2.5 Coding of Data ............................................................38

3.3 Study 3: Training with hand gestures encoding lip aperture and lip rounding for the learning of close-mid and open-mid vowel contrasts ........................................38
3.3.1 Research questions ..........................................................38
3.3.2 Participants .................................................................38
3.3.3 Materials .................................................................39
3.3.4 Experimental procedure ...................................................39
3.3.5 Coding of data ...............................................................39

4. Work plan ............................................................................40

5. References ...........................................................................41

Appendix ..................................................................................49
Abstract

The acquisition of novel speech sounds has been regarded as one of the key challenges in foreign language pronunciation learning. A growing number of studies have shown the positive effects of High Variability Phonetic Training (i.e., the exposure to high-variability auditory stimuli in different phonetic contexts) for L2 novel phoneme acquisition (see Barriuso & Hayes-Harb, 2018, for a review), and especially larger effects have been reported by adding the visual modality from lip movements for certain phonemes, especially when the target phonetic features are visually accessible (e.g., Hazan, Sennema, Iba, & Faulkner, 2005; Inceoglu, 2016). Now the question is whether providing an additional layer of visual modality, namely hand gestures encoding phonetic features, will yield an improvement during phonetic training. Even though previous studies have shown consistent effects of audiovisual training with visuospatial hand gestures encoding prosodic features in the L2, less is known on the effects of hand gestures encoding segmental features. The main goal of this thesis is to further assess the potential role of visuospatial hand gestures encoding phonetic features in the learning of L2 novel phonemic contrasts by focusing on the two issues, namely the appropriateness of the gesture form for the target phonetic feature, and the degree of visual accessibility to the target phonetic feature encoded by hand gesture.

Three phonetic training studies with a pre- and posttest design will be carried out with the aim of assessing the effects of using visuospatial hand gestures on the learning of novel phonemic contrasts, at both the perception and production levels. Study 1 will assess the role of a hand gesture mimicking the burst of air on the perception and production of Mandarin Chinese aspirated and unaspirated consonants (i.e., plosives and affricates) by Catalan naïve learners. Crucially, the hand burst gesture appropriately matches the acoustic difference between aspirated and unaspirated plosives, while this is not the case for the affricates. The results of this study showed that although hand gestures showed limited effects on the perception of aspiration contrasts, they significantly helped the pronunciation of plosives, suggesting that hand gestures are more useful when they appropriately mimic phonetic features. Study 2 will examine whether a hand gesture mimicking barely visually accessible articulatory information, e.g., the lifting of the tip of the tongue, will facilitate the learning of Mandarin Chinese alveolar-retroflex sounds (/ɕ/-/ʂ/; /s/-/ʃ/) by Catalan naïve learners. Study 3 will test whether hand gestures mimicking a visible articulatory
feature, e.g. lip aperture and lip rounding, can help the learning of Catalan open-mid and closed-mid vowels (/ɛ/-/ɛ/, /ɔ/-/ɔ/) by Chinese naïve learners of Catalan. Taken together, studies 2 and 3 will assess the role of the degree of visual accessibility to the target phonetic feature encoded by hand gesture in the learning of L2 phonemic contrasts.

All in all, the studies included in this thesis will assess for the first time the effects of observing a variety of hand gesture forms on the learning of L2 novel phonemic contrasts by focusing on how these gesture forms can optimally encode phonetic information. In this way, the results of this thesis will be able to assess in a more comprehensive way the efficacy of different types of visuospatial gestures in phonetic learning and how they can be implemented to promote optimal learning.

**Keywords:** visuospatial hand gestures, second language, segment learning, speech perception, speech production
Resum

L’adquisició de nous sons de parla ha estat considerat com un dels reptes clau de l’aprenentatge de la pronunciació de llengües estrangeres (L2). Un nombre creixent d’estudis han demostrat els efectes positius dels entrenaments fonètics en els quals hi ha una exposició auditiva dels sons en diferents contextos fonètics (veure Barriuso & Hayes-Harb, 2018, per a una revisió). També s'ha descobert que la presència de la modalitat visual en aquests entrenaments fa que siguin especialment efectius sobretot quan els trets fonètics són accessibles visualment (e.g., Hazan, Sennema, Iba, & Faulkner, 2005; Inceoglu, 2016). L’objectiu principal de la tesi és investigar si el fet de proporcionar una capa addicional de modalitat visual (concretament els gests manuals que codifiquen els trets fonètics de forma visuoespacial), pot desembocar en una millora més gran de l’entrenament fonètic. Tot i que estudis anteriors han mostrat efectes positius de l’entrenament fonètic audiovisual amb gests manuals que codifiquen trets prosòdics, hi ha molt pocs estudis empírics sobre els efectes de l’ús d’aquests gestos de la mà durant els entrenaments fonètics. Aquesta tesi avaluarà el paper dels gests manuals en l’aprenentatge de nous contrastos fonèmics en la L2, centrant-se sobretot en els dos temes següents: l’adекuació de la forma i el moviment del gest per al tret fonètic en qüestió, i el grau d’accessibilitat visual del tret fonètic codificat pel gest manual.

La tesi es composa de tres estudis experimentals d’entrenament fonètic amb dissenys pre- i posttest. L’Estudi 1 avaluarà el paper d’un gest manual que imita el tret fonètic de l’aspiració (la forta sortida de l’aire) en la percepció i la producció de consonants oclusives i africades aspirades i no aspirades xinèss mandarí per part d'estudiants catalans novells. Crucialment, mentre que el gest d’obertura és adequat per als sons oclusius aspirats del xinès, aquest no és el cas per als sons africans. Els resultats d’aquest estudi han mostrat que els gests ajuden significativament a la millora de la pronunciació de les consonants oclusives aspirades i no pas a la pronunciació de les africades, fet que suggereix que els gests manuals són útils quan imiten de forma adequada els trets fonètics. L’Estudi 2 examinarà si un gest manual que imita una informació articulatoria no accessible visualment, per exemple, l’aixecament de la punta de la llengua en els sons retroflexos, facilitarà l’aprenentatge dels sons alveolars-retroflexos del mandarí (/ʈʃ/-/ʈʂ/, /s/-/ʂ/) per part d'estudiants catalans novells. L’Estudi 3 comprovarà si els gests manuals que imiten un tret articulatori visible, per exemple l’obertura i l'arrodoniment dels llavis de
les vocals catalanes semiobertes i semitancades (/e/-/ɛ/, /o/-/ɔ/) poden ajudar a estudiants xinesos principiants de català a pronunciar aquest sons. En conjunt, l'estudi 2 i 3 avaluaran el paper dels diferents graus d'accessibilitat visual del tret fonètic codificat pel gest manual en l'aprenentatge dels contrastos fonèmics d'una L2.

En definitiva, els estudis inclosos en aquesta tesi valoraran per primera vegada els efectes d’observar diverses formes de gestos manuals sobre l’aprenentatge de nous contrastos fonèmics en una L2, centrant-se en com aquestes formes gestuals poden codificar informació fonètica de manera òptima. Així, els resultats d’aquesta tesi podran valorar d’una manera més completa l’eficàcia de diferents tipus de gestos manuals en l’aprenentatge fonètic i com es poden implementar per promoure un aprenentatge òptim.

**Paraules clau:** gestos manuals visuoespacial, segona llengua, aprenentatge de segments, percepció de la parla, producció de la parla
Resumen

La adquisición de nuevos sonidos de habla ha estado considerado como uno de los retos clave del aprendizaje de la pronunciación de lenguas extranjeras (L2). Un número creciente de estudios han demostrado los efectos positivos de los entrenamientos fonéticos en los cuales hay una exposición auditiva de los sonidos en diferentes contextos fonéticos (ver Barriuso & Hayes-Harb, 2018, para una revisión). También se ha descubierto que la presencia de la modalidad visual en estos entrenamientos hace que sean especialmente efectivos sobre todo cuando los rasgos fonéticos son accesibles visualmente (e.g., Hazan, Sennema, Iba, & Faulkner, 2005; Inceoglu, 2016). El objetivo principal de la tesis es investigar si el hecho de proporcionar una capa adicional de modalidad visual (concretamente los gestos manuales que codifican los rasgos fonéticos de forma visuoespacial), puede desembocar en una mejora más grande del entrenamiento fonético. Aunque los estudios anteriores han mostrado efectos positivos del entrenamiento fonético audiovisual con gesto manuales que codifican rasgos prosódicos, hay muy pocos estudios empíricos sobre los efectos del uso de estos gestos de la mano durante los entrenamientos fonéticos. Esta tesis evaluará el papel de los gestos manuales en el aprendizaje de nuevos contrastes fonémicos en la L2, centrándose sobre todo en los dos temas siguientes: la adecuación de la forma y el movimiento del gesto para el rasgo fonético en cuestión, y el grado de accesibilidad visual del rasgo fonético codificado por el gesto manual. La tesis se compone de tres estudios experimentales de entrenamiento fonético con diseños pre- y posttest. El Estudio 1 evaluará el papel de un gesto manual que imita el rasgo fonético de la aspiración (la fuerte salida del aire) en la percepción y la producción de consonantes oclusivas y africadas aspiradas y no aspiradas del chino mandarín por parte de estudiantes catalanes noveles. Crucialmente, mientras que el gesto de apertura es adecuado para los sonidos oclusivos aspirados del chino, este no es el caso para los sonidos africados. Los resultados de este estudio han mostrado que los gestos ayudan significativamente en la mejora de la pronunciación de las consonantes oclusivas aspiradas y no en la pronunciación de las africadas, hecho que sugiere que los gestos manuales son útiles cuando imitan de forma adecuada los rasgos fonéticos.

El Estudio 2 examinará si un gesto manual que imita una información articulatoria no accesible visualmente, por ejemplo, el levantamiento de la punta de la lengua en los sonidos retroflexos, facilitará el aprendizaje de sonidos alveolares-retroflexos del mandarín (/ths/-
/ts/, /s/-/ʃ/) por parte de estudiantes catalanes noveles. El Estudio 3 comprobará si los gestos manuales que imitan un rasgo articulatorio visible, por ejemplo, la apertura y el redondeo de los labios de las vocales catalanas semiabiertas y semicerradas (/e/-/e/, /o/-/ɔ/) pueden ayudar a estudiantes chinos principiantes de catalán a pronunciar estos sonidos. En conjunto, los estudios 2 y 3 evaluarán el papel de los distintos grados de accesibilidad visual del rasgo fonético codificado por el gesto manual en el aprendizaje de los contrastes fonémicos de una L2.

En definitiva, los estudios incluidos en esta tesis valorarán por primera vez los efectos de observar distintas formas de gestos manuales sobre el aprendizaje de nuevos contrastes fonémicos en una L2, centrándose en cómo estas formas gestuales pueden codificar información fonética de manera óptima. Así, los resultados de esta tesis podrán valorar de una forma más completa la eficacia de distintos tipos de gestos manuales en el aprendizaje fonético y cómo se pueden implementar para promover un aprendizaje óptimo.

**Palabras clave:** gestos manuales visuoespaciales, segunda lengua, aprendizaje de segmentos, percepción del habla, producción del habla
1. Introduction

1.1 Object of analysis

The research in this PhD thesis will broadly assess the effects of audiovisual phonetic training with visuospatial hand gestures encoding phonetic features of novel phonemic contrasts on the learning of these contrasts by naïve L2 learners. Crucially, the studies included in this PhD thesis will analyze to what extent the correspondence patterns between gesture form and the encoded phonetic features will influence the training effect.

Previous studies have shown that training with visuospatial hand gestures favors the learning of novel suprasegmental features, such as lexical tones, intonation, speech prominence and word stress. Yet little is known about the potential effects of visuospatial hand gestures encoding phonetic features on the learning of novel phonemic contrasts. Importantly, the correspondence between gesture form and the encoded learning features may influence the learning effect, as it has been shown in other linguistic contexts (e.g., Kelly, McDevitt, & Esch, 2009). Study 1 will assess whether hand gestures matching or mismatching the phonetic properties of novel phonemes will yield different learning effect. Moreover, results obtained in audiovisual phonetic training studies suggest that benefits from audiovisual training (AV benefit) may largely depend on the degree of visual accessibility of the phonetic features of the target sounds (e.g. Hazan et al., 2005). Studies 2 and 3 will assess whether hand gestures encoding different levels of visually accessible phonetic features (visible feature of lip aperture and lip rounding versus invisible feature of tongue tip lifting) will have different training effects.

Three training studies with a pre- and posttest design will be conducted to test the abovementioned questions. The first study will evaluate the effects of a hand burst gesture which visually depicts aspiration on the learning of nonnative aspiration contrasts by naïve Catalan learners of Chinese. Importantly, the study will assess whether the presence of a mismatch between the hand gesture and the target articulatory features influences phonological learning. Crucially, while the hand burst gesture used in the study closely matches the properties of the aspirated plosives, this gesture mismatches with the affricates. The second study will assess whether training with a palm-down finger-lift gesture which visually encodes the lifting of the tip of the tongue for retroflex consonants (a less accessible phonetic feature) will help Catalan speakers learn Mandarin alveolar-retroflex
contrasts (\(\text{tʃ}/-\text{ʃ}/, \text{s}/-\text{ʂ}/\)). The third study will assess whether training with hand gestures visually encoding lip aperture and lip rounding (a visible phonetic feature), will facilitate Chinese naïve learners of Catalan to learn Catalan closed-mid and open-mid vowel contrasts (/e/-ɛ/, /o/-ɔ/). Taken together, the second and third studies will examine whether hand gestures encoding different degrees of visually accessible phonetic information will yield different training effects on L2 segment learning.

Importantly, the three studies will include an assessment of both perceptual phoneme identification tasks and production tasks in the pre- and posttest sessions. This will provide more comprehensive information concerning the learning patterns of L2 phonological features.

All in all, the studies included in this thesis will be the first to control and vary in a systematic way the correspondence patterns between gesture form and phonetic features of target L2 phonemic contrasts. In this way, the results of this thesis will be able to assess in a more comprehensive way the efficacy of different types of visuospatial gestures in phonetic learning and how they can be implemented to promote optimal learning.

1.2 Literature Review

1.2.1 L2 pronunciation learning and teaching

In the foreign language classroom, pronunciation tends to receive less attention compared with other linguistic components, such as syntax and morphology (Deng et al., 2009). Some studies suggest that the reason why teachers do not provide enough pronunciation training might be due to the lack of formal training in teaching pronunciation, the lack of teaching and learning materials, or the lack of confidence in giving pronunciation instruction (e.g., Breitkreutz, Derwing, & Rossiter, 2001; Macdonald, 2002). Breitkreutz and colleagues (2001) conducted a survey with 67 L2 English teaching programs. Although an overwhelming majority of the respondents reported that both suprasegmental and segmental features were important aspects of L2 pronunciation, interestingly, they believed that pronunciation instruction on difficult segments was the best strategy to improve L2 learners’ pronunciation. More recently, Darcy, Ewert and Lidster (2012) provided some complementary reasons for why L2 English teachers did not give enough pronunciation instruction to students. Fourteen teachers answered 15 questions in four categories and the analysis of each category showed that: (a) 92% of the
teachers considered pronunciation instruction as important for students; (b) the first five most important pronunciation elements they thought were intonation/stress, rhythm, perception ability, clarity of individual sounds, and specific vowels; (c) the instruction of specific pronunciation elements should be taught at different proficiency levels, and more specifically, learners with low levels of L2 proficiency should receive more instruction related to segmentals, while those with high levels of proficiency should be taught more on suprasegmentals; (d) while 71% of the teachers reported that they did not teach pronunciation at all, 28% of them would provide a diagnostic task before the course in order to decide which phonological features should be addressed in class; and finally 14% reported that they would do an evaluation with a pronunciation rubric after the course. Overall these results suggest the difficulty for teachers to manage pronunciation instruction in their classrooms. Some of the reasons that were identified by these same teachers were lack of available time (43%), lack of training on how to teach pronunciation (25%), and the need for more guidance and support from the institution (18%).

Apart from L2 instructors, students are also concerned with their pronunciation accuracy (e.g., Derwing, 2003; Rajadurai, 2001). In Derwing (2003), 100 L2 English learners were interviewed on their attitude towards accent and their pronunciation problems. Results showed that half of the students considered pronunciation as an important factor in communication in the second language. Crucially, among all the pronunciation problems faced by these students, the problems with segmental features ranked top, with 26 students identifying the difficulty of pronouncing English /θ/ and /ð/ sounds, 12 of them highlighting the distinction between /l/ and /r/, and 28 of them mentioning other consonants and vowels. Similarly, Rajadurai (2001) investigated 74 L2 English learners’ views on the teaching of pronunciation. Results showed that most of students considered pronunciation training as indispensable and helpful to enhance the ability in perceiving auditory discrimination, as well as to produce speech more clearly and with more confidence. Interestingly, 91% of them reported that training in segmentals (i.e., consonants and vowels) was more helpful than training in suprasegmentals (e.g., stress, weak forms, intonation, etc.), as they felt that training in speech sounds could lead them to improve their L2 pronunciation.

All in all, it is clear that pronunciation instruction remains a challenge in L2 teaching and learning. Also, there is a need for more empirical research into how to improve pronunciation skills. Following this line of research, numerous empirical studies have assessed different methods to teach L2 pronunciation in an effective way. Among these studies, L2 segment learning has received considerable attention by researchers. For adult
learners, the acquisition of L2 phonemic contrasts has been regarded as one of the major difficulties when learning pronunciation, in terms of both the creation of accurate mental representations of the phonemes and the ability to perceive and pronounce them. The main objective of this PhD thesis is to examine the efficiency of one particular training method for segmental pronunciation, specifically the use of visuospatial hand gestures. Next section summarizes an L2 phonetic training approach on the learning of novel phonemic contrasts, namely High Variability Phonetic Training.

1.2.2 High Variability Phonetic Training (HVPT)

Over the last decades, many studies have investigated different training methods to teach novel L2 phonemes. High Variability Phonetic Training (HVPT) is one of the most frequently used approaches in L2 phonetic training and it has been shown to be effective in many empirical studies (see Barriuso & Hayes-Harb, 2018; Thomson, 2008, for a review). It exposes learners to a large amount of auditory input from a variety of speakers and phonetic environments, instead of limiting the input to the production of a single speaker or one consistent linguistic context for target sounds.

This approach was first developed by Logan, Lively and Pisoni (1991) with the focus on the perception of English /l/-/r/ contrast. In this study, 6 Japanese listeners undertook a 3-week training program. The training stimuli included a total of 68 minimal pairs of words produced by 5 English native speakers and containing the target English /l/ and /r/ sounds contrasting in different syllable positions (i.e., syllable initial, medial, or final positions). During the training phase, participants were asked to identify the stimulus they heard and were given immediate feedback after each trial as to whether or not their response was correct. Their perception accuracy of the target contrast was tested before and after the training phase with an identification task. The results showed that Japanese listeners were significantly better at identifying the target contrast after training, suggesting that the HVPT method could boost the perception of difficult L2 phonemic contrasts.

Moreover, HVPT also has been shown to help learners to generalize the learning effect (Cebrian & Carlet, 2014; Lively et al., 1993; Logan et al., 1991). If the generalization is observed, robust learning is believed to be achieved, as generalization is understood as the ability to transfer learned knowledge to several dimensions. Indeed, Logan et al. (1991) tested 3 of the participants after the posttest with new testing stimuli produced by a familiar talker (one of the speaker in the training session) and another set of novel stimuli produced
by a new talker. Results showed that the participants could generalize phonological learning to novel speech produced by a familiar speaker and also to novel stimuli produced by a novel speaker, which was confirmed by a follow-up study (Lively et al., 1993). Interestingly, in the study of Cebrian and Carlet (2014), the effects of short-term HVPT (four 45-minute sessions) was tested on the perception of English consonantantal and vocalic contrasts by Catalan/Spanish bilingual learners of English. Results showed that although HVPT was only effective for some vowels and consonants, the generalization of improvement on perception was detected, suggesting that robust learning could occur even after a short period of HVPT.

Although the canonical HVPT is a perception-only approach that provides learners with auditory exposure, it has been shown to be effective in improving L2 production (e.g., Bradlow, Pisoni, Akahane-Yamada, & Tohkura, 1997). In Bradlow et al. (1997), the effects of HVPT on speech production were examined by testing the pronunciation of words containing /l/ and /r/ by 11 Japanese learners of English. The results showed that HVPT facilitated the production of L2 phonemic contrasts, suggesting that perceptual gains during HVPT could transfer to the production domain.

HVPT could facilitate the perception and production of L2 speech sounds, generalize the improvement, and also encourage a retention of learning effect over the long term (e.g., Bradlow, Akahane-yamada, Pisoni, & Tohkura, 1999; Lively, Pisoni, Yamada, Tohkura, & Yamada, 1994). Moreover, this approach has been shown to be effective in the learning of various L2 speech sounds by learners of different L1 backgrounds, such as the perception and production of English mid and low vowels by Japanese learners (Lambacher et al., 2005), as well as the perception of 14 British English vowels by Spanish and German learners (Iverson & Evans, 2009), among many others.

Overall, this section reviews the positive effects of HVPT. As suggested by Aliaga-Garcia and Mora (2008), the effectiveness of phonetic training does not rely on the amount of training time, the quality of training matters. HVPT provides a quantity and diversity of auditory input to L2 learners which helps them create new phonemic categories in mental representation and improve pronunciation performance.

Despite the fact that HVPT is largely successful in phonetic training, some work has shown how adding a visual modality to this paradigm seems to generate additional learning effects. Indeed, audiovisual phonetic training combines canonical HVPT with articulatory information in the visual modality. The following section introduces this training approach and summarizes the effect of this training paradigm on L2 segment learning.
1.2.3 Audiovisual phonetic training

Research in the last few decades has shown that language processing and language learning is multimodal in essence. Humans perceive speech through both auditory and visual information, and the two dimensions interact. First, speech perception can be influenced by visual input. The well-known McGurk effect (McGurk & MacDonald, 1976) has shown that speech perception may be strongly affected by visual information from speech articulators. When seeing the visual information of /ga/ from lip movements and hearing the auditory stimuli of /ba/, the conflict between the two modalities may result in the perception of a /da/ sound, suggesting that the visual modality related to the speech sound is integrated with the auditory modality before the sound is categorized. Many other studies have shown that the integration of auditory and visual information enhances listeners’ speech perception in both noisy (Grant & Seitz, 2000) and normal environments (Arnold & Hill, 2001; also see Summerfield, 1992, for a review) and can improve speech recognition for hearing-impaired listeners (Grant et al., 1998) and nonnative listeners (Navarra & Soto-Faraco, 2007).

Building on the important role of the visual modality in speech perception, several studies have investigated whether the addition of the visual modality to the HVPT paradigm would further enhance the perception of L2 phonemic contrasts and lead to better improvement in pronunciation (Hardison, 2003; Hazan et al., 2005). In comparison with HVPT, audiovisual training allows learners not only to hear the target sounds, but also to see articulatory information while the speakers are producing the sounds.

In Hardison (2003), 16 Japanese native speakers and 8 Korean native speakers were trained in 15 30-minute sessions on English /r/-/l/ contrast in one of the following two conditions. In the Audio-Only condition, participants could only get access to auditory input produced by 5 English native speakers. In the Audiovisual condition, participants watched a training video in which both the auditory and visual information were provided. A control group of 8 Japanese and 8 Korean speakers received the testing but no training. Their perception performance was tested before and after training by an identification task. In the identification task, the testing materials for the Audio-only group were presented auditorily, while materials for the Audiovisual group were presented in three different conditions: (a) Audio-only, in which the participants could only hear the auditory stimuli; (b) Audiovisual, in which the participants could hear the stimuli and see the lip movements; and (c) Video-only, in which the participants could only see the lip movements. Results
revealed that audiovisual training yielded significantly greater improvement on perceptual accuracy than training with pure auditory information for both Japanese and Korean participants.

However, the extent to which the visual modality can facilitate the perception of L2 phonemic contrasts is shown to be dependent on the visual accessibility of the articulatory features to the target contrasts (Hazan et al., 2005). In this study, Hazan and colleagues investigated the effect of audiovisual perceptual training on the perception of two pairs of English consonantal contrasts which have different degrees of visual accessibility to their articulatory information: a visually salient phonetic contrast between labial (/b/ and /p/) and labiodental (/v/) sounds (Experiment 1), and a less visually accessible contrast between /r/ and /l/ (Experiment 2). One hundred and one participants (39 in Experiment 1; 62 in Experiment 2) were trained in one of the following two conditions: Audiovisual or Audio-only condition. The results showed that audiovisual training was more effective in improving the perception of the labial/labiodental contrast compared to auditory training alone, whereas there was no difference between the two training methods for the /r/-/l/ contrast. Since the distinction of articulatory information for /r/-/l/ contrast is less visually accessible than the labial/labiodental contrast, these results suggest that the amount of benefits from the integration of auditory and visual modality (known as, AV benefit) seems to depend upon whether the articulatory movements of the L2 phoneme contrasts involve visually accessible information.

Interestingly, improvement in sound perception also led to improvement in pronunciation accuracy. In Hardison (2003), participants’ production ability was tested through a reading task, in which participants were asked to read a word containing a target phonemic contrast presented on a card. The accuracy of target production was rated by 5 English native speakers using a 7-point scale. The results showed that participants’ production performance improved significantly after both audiovisual and audio-only training, revealing a transfer of the improvement in perception to pronunciation. Similarly, in Hazan et al. (2005), participants’ productions of the /r/-/l/ contrasts were tested through a reading task and were measured through an identification task and a rating task performed by 12 English native speakers. The results showed that in general, audiovisual training led to a greater improvement in pronunciation of the /r/-/l/ contrast compared to the audio-only training. A key finding of this study was thus that even though adding the visual modality was not more helpful in the perception of the less visually accessible contrast between /r/ and /l/, audiovisual training seemed to have a more beneficial effect on pronunciation. This
finding was further confirmed by Inceoglu's (2016) study, in which 60 English learners of French were trained in the learning of three French nasal vowels. Participants were divided into three training modalities: audiovisual training, audio-only training or no training. The participants’ perception and production of the target vowels was assessed with an identification task and an imitation task before and after training. The results showed that audiovisual training was not more effective than two conditions on improving the perception of the French nasal vowels, but led to a significantly greater improvement in the production performance. Overall, these findings suggest that audiovisual training (i.e., a phonetic training which combines auditory and visual modalities) can have stronger beneficial effects in the learning of L2 phonemic contrasts than audio-only phonetic training.

To sum up, the studies summarized in this section show phonemic detection accuracy and the pronunciation of nonnative phonemic contrasts can be enhanced via audiovisual perceptual training, and particularly that audiovisual training is more effective than auditory training when the phonemic contrasts are sufficiently salient visually. Apart from the benefits of the visual modality from articulatory information, hand gestures could also be integrated in the phonetic training as an additional visual modality. In the next section, we summarize the effects of hand gestures on L2 language learning.

1.2.4 Effects of hand gestures on language learning

1.2.4.1 The benefits of hand gesture for L2 acquisition

Hand gesture and speech have been considered as an integrated system as they appear to be governed by a single process. McNeill states that “… hand gestures and speech are closely linked in meaning, function, and time; they share meanings, roles, and common fate” (McNeill, 1992, p. 218). According to McNeill (1992), co-speech gestures and speech are integrated in three complementary ways: (a) gestures can convey the meaning of speech, (b) gestures can show same pragmatic function as speech, and (c) gestures are phonologically integrated with speech.

Apart from the integration between hand gestures and speech, gestures also play an important role in language processing and language learning. For example, referential gestures have been shown to facilitate the recall of information in a first language (L1) by adults (e.g., So, Sim Chen-Hui, & Low Wei-Shan, 2012) and children (e.g., Cook, Mitchell,
& Goldin-Meadow, 2008; So et al., 2012); referential gestures also boost speech comprehension in an L1 (Cocks, Morgan, & Kita, 2011; also see Hostetter, 2011, for a review). As mentioned in Hostetter (2011, p. 299), “even when a gesture is redundant with the accompanying speech, listeners may attend to the gesture as an alternative source of information when they cannot understand the speech”. This suggests that even though hand gestures provide a nonverbal means of conveying same information as speech, they can help listeners to better understand speech.

McCafferty (1998) stated that “Gestures and other nonverbal forms of communication have been considered potentially important for some time, however, as yet, their connection to second language learning largely remains to be elucidated” (p.94). Gullberg (2006) outlined several reasons for studying the role of hand gestures in L2 acquisition. First, the author claimed that hand gestures can be treated as an essential part of language acquisition since gestures are a cross-cultural phenomenon. Second, considering the close connection between gesture, language and speech, the development of hand gestures by learners can offer insights into their learning processes. Moreover, as a visual input to both learners and interlocutors, hand gestures may lend a hand in comprehension and learning.

L2 classroom research shows that teachers use a variety of gestures to support L2 learning. Allen (2000) observed the teaching behavior of a female teacher in a classroom setting. The researcher noted that the teacher used various hand gestures for different instructional purpose and students reported that hand gestures helped them to maintain the focus of attention, making the class more interesting and appealing by setting “a more relaxed and casual atmosphere” (Allen, 2000, p. 169), as well helping students to better understand what the teacher was trying to explain and keep them interested in the learning process. Sime (2006) interviewed 22 L2 English learners regarding the perception of functions of hand gestures used by teachers in the classroom. The author pointed out the following three main functions of observing hand gestures by students in the L2 classroom: (a) cognitive, i.e., hand gestures facilitate the learning process; (b) emotional, i.e., hand gestures serve as a communicative device of teachers’ emotion; and (c) organizational, hand gestures are employed in order to manage classroom interaction. Sime (2008) again considered student perception of teachers’ gestures by examining five L2 English classrooms. All of the learners reported that gestures were helpful in relating meaning and enhancing comprehension. They also reported that gestures contributed positively to the process of classroom interaction.

Apart from the pedagogical use of hand gestures in L2 classrooms, a large body of
empirical studies has assessed the role of different types of co-speech gestures in different learning contexts. According to McNeill (1992), hand gestures can be broadly divided into two groups: referential and non-referential gestures. Referential gestures refer to explicit semantic meaning conveyed in speech, which consist of iconic (i.e., hand gestures mimicking certain entity or action), metaphoric (i.e., hand gestures representing abstract concepts), and deictic gestures (i.e., hand gestures pointing to certain object). Non-referential gestures are hand gestures that do not convey meaning referring to speech, such as beat gestures which are typically associated with prosodically prominent positions in speech. In recent years, studies have shown that referential gestures could facilitate novel word memorization (e.g., Kelly et al., 2009; Macedonia, Müller, & Friederici, 2011; Quinn-Alley, 1995; Tellier, 2008), that non-referential beat gestures have also been shown to favor L2 word learning (Kushch et al., 2018), and that hand gestures help L2 learners to better comprehend nonnative speech (Dahl & Ludvigsen, 2014), especially when their language proficiency is low (Sueyoshi & Hardison, 2005).

In the following section, we will focus on the beneficial role of hand gestures in L2 pronunciation learning, at both the prosodic (1.2.4.2) and the phonetic (1.2.4.3) levels.

1.2.4.2 The benefits of visuospatial hand gesture for L2 suprasegmental learning

In the field of L2 pronunciation instruction, teachers use a variety of hand gestures to help students achieve L2 target phonological features of pronunciation (e.g., Hudson, 2011; Smotrova, 2017; Zhang, 2002). Specifically, teachers use hand gestures visually mimicking certain segmental or suprasegmental features in space to support their explanation and instruction of target features, and students observe and sometimes imitate the hand gestures. For example, in Smotrova (2017), the teacher nodded the head and tapped the fingers to help students realize the number of syllables when instructing syllabification, she made an consistent upward body movement to teach word stress, and she gestured upward and downward with head tilts and even moved the whole body to teach rhythm. Zhang (2002) proposed a set of hand gestures which encode different phonological features for teaching difficult speech sounds of Mandarin Chinese, for example, a fist-to-open hand gesture for aspirated consonants, a fist gesture for rounded vowels, etc. Hudson (2011) observed a university English as second language pronunciation course in which teachers integrated gestures into pronunciation teaching. Apart from using gestures to indicate syllable stress and intonation, a large number of
gesture uses were observed for teaching the phonemes, such as drawing a horizontal line with the index finger or extending both arms outward to teach sounds with longer duration, moving closed fingers of both hands away from middle of mouth to the sides and back to the middle to illustrate the mouth movement of the sound /æ/, doing an “o” shape gesture to show the lip roundness of the sound /u/ (see Appendix A).

The visuospatial hand gestures refer to a variety of hand gestures encoding certain phonological properties. According to the specific phonetic and prosodic features that hand gestures encode, the type of visuospatial gestures can be classified as (a) pitch gestures (i.e., hand gestures which mimic pitch contours in space), (b) beat gestures (i.e., hand gesture which highlight prosodic prominence through up-and-down rhythmic movements), (c) durational hand gestures (i.e., hand gestures showing duration properties of phonemic length contrast), and (d) hand gestures encoding phonetic features (i.e., hand gestures highlighting articulatory or acoustic features of phonemes).

A large body of empirical studies have examined the effectiveness of hand gestures on L2 suprasegmental features, such as tone, intonation, rhythm and durational properties (e.g., Baills, Suárez-González, González-Fuente, & Prieto, 2019, among others), whereas few studies have been conducted on the learning of L2 phonemic contrasts (Amand & Touhami, 2016; Hoetjes et al., 2019).

Pitch gestures (i.e., hand movements which mimic pitch contours in space) have been shown to enhance L2 lexical tone perception (e.g., Baills et al., 2019; Morett & Chang, 2015), and L2 intonation production (Yuan et al., 2019). In Baills et al. (2019), 49 adult Catalan native speakers without prior knowledge of Mandarin Chinese were audiovisually trained to perceive the four Chinese tones either with or without simultaneous pitch gestures (see Figure 1). The results showed that participants who observed pitch gestures improved more on their perception of lexical tones at posttest than participants who were only exposed to audiovisual speech.
**Figure 1.** Pitch gestures of the four Mandarin tones as well their corresponding waveforms and pitch tracks. Reproduced from Baills et al. (2019).

Yuan et al. (2019) showed that the use of pitch gestures boosted L2 beginners’ pronunciation of L2 nuclear pitch configurations. A total of 64 Chinese undergraduate students of Castilian Spanish were trained to pronounce three different types of Spanish phrase-final intonational pitch contours, either with or without pitch gestures. The results of a sentence imitation task showed that observing pitch gestures improved participants’ pronunciation of these pitch contours significantly more than not observing pitch gestures.

The beneficial role of beat gestures has been shown by Gluhareva and Prieto (2017). In this study, 20 Catalan learners of English were trained with easy and difficult prompts in English, either with or without accompanying beat gestures. The results showed that participants’ pronunciation of the difficult English prompts improved significantly more when they were trained with beat gestures than when they were not.

In addition, a set of studies have assessed the role of hand gestures in the learning of Japanese vowel length contrasts (e.g., Hirata & Kelly, 2010; Hirata, Kelly, Huang, & Manansala, 2014; Kelly, Bailey, & Hirata, 2017; Li, Baills, & Prieto, 2020). Hirata and Kelly (2010) trained 60 English native speakers to distinguish five pairs of Japanese vowel length contrasts via different modal inputs: (a) audio-only, (b) audio with mouth movements, (c) audio with hand movements, or (d) audio with both mouth and hand movements. The results of a vowel identification task revealed that participants improved their perception of long and short vowel contrasts in general, but audio with mouth movements yielded significantly larger improvements than the audio-only condition and adding gestures was not more helpful than exposure to the audio-only condition. Follow-up studies found that the use of hand gestures may not help the learning of vowel durational contrasts (Hirata et al., 2014; Kelly et al., 2017). The authors argued that in this case, multimodal input might not have been effective due to cognitive overload and visual distraction, as hand gestures may have distracted participants’ attention from the speech signal or mouth and lip movements. The authors also speculated that they “may have chosen a wrong type of gesture to distinguish long and short vowels in language perception” (Hirata & Kelly, 2010, p. 305). Crucially, this choice might have contributed to the negative results obtained, as the use of a beat gesture for the short vowel and the hand sweep gesture for the long vowel may not have been totally appropriate to represent the durational contrasts given behavioral evidence that the mental representation of temporal duration is
visually associated with horizontal displacement (Casasanto & Boroditsky, 2008). Thus, Li et al. (2020) examined the use of durational hand gesture (see Figure 2) on the learning of these vowels by Catalan speakers. They found that even though this type of gesture had limited effects on improving the perception of durational contrasts, it positively influenced the pronunciation of words carrying durational contrasts by naïve participants.

Figure 2. Durational hand gestures of Japanese short and long vowels. Reproduced from Li et al. (2020).

Interestingly, the Verbotonal Method (VTM) which trains prosodic features such as rhythm, accentuation and intonation, has been shown to be helpful in L2 pronunciation (e.g., Alazard-Guiu, Santiago, & Mairano, 2018; Alazard, Astésano, & Billières, 2012). Crucially, VTM encourages the use of body movements, in which hand gestures have been widely used for phonetic correction at both segmental and suprasegmental levels (Guberina, 1971, 1981). In the study by Alazard-Guiu et al. (2018), 8 English beginners of French were trained on reading skills with either of the two methods: (a) the Articulatory Method, in which the explicit articulatory descriptions of different L2 sounds were provided, and (b) the VTM, which combined rhythmic patterns and hand gestures (e.g., a rising hand movement for salient syllables). The results showed a significant improvement on F3 value of /a/ sound regardless of the training methods. However, because of the limited number of participants and the lack of significance between the two methods, more work should be done to provide more evidence on the beneficial effects of L2 prosodic training with hand gestures on L2 segment learning.

1.2.4.3 The benefits of visuospatial hand gesture for L2 phonetic learning

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

In the context of the learning of L2 phonemes, previous studies have shown beneficial effects of audiovisual training on both perception and production. Whether adding hand gestures encoding phonetic features as an additional visual cue to the audiovisual training could yield additional improvements still remains unclear. Despite the wide use of hand gestures cueing phonemes in L2 classrooms (e.g., Hudson, 2011; Zhang, 2002), Lan and Wu (2013) showed that a form-focused instruction with the help of hand gestures mimicking the tongue position (see Appendix A) was effective to improve the pronunciation of English /r/ by native Mandarin speakers. However, only two empirical studies have assessed the effects of hand gestures cueing articulatory features on L2 segment learning (Amand & Touhami, 2016; Hoetjes et al., 2019). Appendix 1 shows a
summary table of the use of hand gestures cueing phonemes in L2 pronunciation learning and in clinical field.

Amand and Touhami (2016) showed the positive effects of observing gestures mimicking the released and unreleased English stops /p/, /t/ and /k/ on their pronunciation by 16 French learners of English. The learners were divided into two training groups. In the gesture group, participants watched a video where the released and unreleased contrasts were explained explicitly and the instructor produced words and sentences containing the target phonemes while using a fist closed hand gesture to represent the unlocked stops and an open palm gesture to indicate the burst of aspiration of the released stops (see Figure 3). Then, participants in this group were asked to read the same speech sample used by the instructor. In the non-gesture group, participants only listened to the audio recordings of the instructor pronouncing the same speech sample and then repeated it after counting to five. Participants’ production was tested by means of a reading task which was undertaken before and after training. An acoustic analysis of the target stop consonants in the reading task revealed that training with gestures mimicking phonetic features yielded significant improvement in the pronunciation of L2 English released and unreleased stops.

Figure 3. Hand gestures mimicking the unreleased (left panel; fist closed) and released (right panel; open palm) English stops. Reproduced from Amand & Touhami (2016).

Recently, Hoetjes and colleagues (2019) trained 51 Dutch native speakers to imitate two Spanish phonemes /θ/ (nonnative) and /u/ (native but with a different phoneme-to-grapheme conversion) in four different conditions: (a) in the audio-only condition, participants listened to the training sentences containing the target phonemes; (b) in the audiovisual condition, participants watched a video in which they were able to see the lip movements of the instructor while listening to the sentences; (c) in the audiovisual condition with pointing gestures, participants watched the instructor performed pointing
gestures (pointing a finger towards the mouth) when pronouncing each target phoneme of the sentences; and (d) in the audiovisual condition with gestures mimicking articulatory information, participants watched a video in which the instructor used her/his right hand to mimic the articulation of the target phoneme when pronouncing it such that the palm and the fingers formed an “o” shape in order to indicate the rounding of the lips for the phoneme /u/, and fingers were extended forward to indicate the tongue position for the /θ/ (see Figure 4). Participants were asked to read 16 four-word Spanish sentences containing the target phonemes before and after the training. Their production was recorded and then subjected to two complementary analyses. In the first, two phoneticians performed a phonetic classification of consonants and vowels in the recordings. In the second, a set of 46 native Spanish speakers rated each recording for perceptual accentedness and comprehensibility on a scale from 1 to 7. The results from both analyses showed that observing pointing gestures had a beneficial effect on the learning of the two phonemes, while observing gestures mimicking articulatory information helped the learning of /u/ but impaired the learning of /θ/. The authors argued that since /u/ already existed in the Dutch phoneme inventory, this sound was relatively easy for Dutch speakers to produce. However, with regard to the impact of training on production of the nonnative phoneme /θ/, only pointing gestures had positive effects, suggesting that different types of gestures have different effects depending on the L2 phoneme to be learned.

Figure 4. Gestures mimicking articulatory information for Spanish phonemes /u/ (left panel) and /θ/ (right panel). Reproduced from Hoetjes et al. (2019).

To sum up, gestures are effective in boosting language processing and learning, not only in L1 but also in the L2. One important theoretical paradigm supporting findings is the Embodied Cognition paradigm, which suggests that cognition is not independent from body and they are related to each other, as cognition is grounded in the sensory modality of brain and in bodily actions (Ionescu & Vasc, 2014). The implication of Embodied Cognition on education has been discussed by some researchers. For example, Shapiro and Stolz (2019) stated four educational implications of Embodied Cognition by focusing on
the use of hand gestures: (a) teachers could increase their gesture use when giving instructions; (b) teachers could encourage students to actively produce hand gestures; (c) teachers should choose different types of gestures for distinct teaching purpose; and (d) teachers should keep updated with the effectiveness of hand gestures validated by research in this field. Since previous studies have shown inconsistent results regarding audiovisual phonetic training with hand gestures encoding phonetic features on the production of L2 phonemes, the present PhD dissertation aims at clarifying the potential effects of this training method on the production as well as on the perception of L2 phonemic contrasts, thus it clearly falls within the framework of Embodied Cognition and embodied learning.

Crucially, the negative effect of training with hand gestures mimicking the tongue shape on the learning of the Spanish consonant /θ/ was detected by Hoetjes et al. (2019). In our view, these findings are due to an inappropriate gesture form, as it may have failed to convey important information about how to articulate this sound such as the fact that the tip of the tongue should be placed between the teeth, and the jaw should be in a fairly open position. The importance of the correspondence between the form of the gesture and the target linguistic property it is representing has been assessed by several studies. First, speech perception may be influenced according to the matching/mismatching of the gesture form (Hannah et al., 2017; Kelly et al., 2017). Hannah and colleagues (2017) found that pitch gestures could facilitate L2 lexical tones perception when the gestures matched the properties of the lexical tones (e.g., a rising gesture for a rising tone), whereas a mismatch between gesture form and pitch movement (e.g., a rising gesture for a falling tone) disrupted perception. Similarly, Kelly et al. (2017) found that congruous pitch gestures facilitated the perception of intonation in L2 Japanese, and that congruous hand gestures mimicking durational features tended to help the processing of vowel length contrast. Second, the use of appropriate hand gesture could boost L2 word learning (Kelly et al., 2009; Macedonia et al., 2011). Kelly and colleagues (2009) trained naïve Japanese learners on 12 common Japanese words in the following four within-subject conditions: (a) speech, (b) speech with congruent iconic gesture (e.g., the gesture of drinking matches the word “drink”), (c) speech with incongruent iconic gesture (e.g., the gesture of washing mismatches the word “drink”), and (d) repeated speech. Results showed that better word memorization was found with words trained with congruent gesture whereas a mismatching gesture did not facilitate L2 word learning. Thus, the effect of training with hand gestures could be influenced by the gesture form in L2 word learning, yet little is known about this effect in the context of L2 pronunciation learning. To fill this gap, Study
1 of this PhD project will assess to what extent the form of the gesture being appropriate to the target phonetic feature will influence the learning outcome by means of examining a fist-to-open-hand gesture on the learning of Mandarin aspiration contrasts.

Mandarin has a total of six pairs of consonants which have traditionally been described as contrasting in aspiration. They are three pairs of plosives (voiceless unaspirated sounds: bilabial /p/, dental /t/ and velar /k/, with their aspirated counterparts /pʰ/, /tʰ/ and /kʰ/) and three pairs of affricates (voiceless unaspirated sounds: alveolar /t̚s/, alveolopalatal /t̚ɕ/ and retroflex /ʈ̚ʂ/, along with their aspirated counterparts /t̚sʰ/, /t̚ɕʰ/ and /ʈ̚ʂʰ/) (Duanmu, 2007). Importantly, though Mandarin plosives and affricates have been traditionally described as having an “aspiration” contrast, the two types of consonants differ in the strength of the air burst and the duration of the frication phase (Laver, 1994). Auditorily, aspirated plosives are characterized by a very salient audible air burst that is due to the changes in air pressure and the sudden release of air. By contrast, aspirated affricates are not perceived to start with an air burst, as the longer release time allows the compressed air to be released over a longer frication period. Catalan plosives and affricates contrast in voicing features, and it can be said that the aspiration feature present in Mandarin plosives and affricates does not exist in the Catalan sound counterparts. Also, the acquisition of the aspiration feature is considered to be one of the main difficulties experienced by speakers of Romance languages when learning Mandarin (Chen et al., 2013). Therefore, both the perception and production of Mandarin aspirated consonants can be considered as difficult to learn for Catalan speakers. As such, Mandarin Chinese is selected as the target L2, and L1 Catalan speakers will be recruited as our subjects. We adopted a fist-to-open-hand gesture (see Figure 5) from Amand and Touhami (2016) and Zhang (2002) to simulate the extra burst of air which refers to the aspiration feature of Mandarin aspirated plosives as they are perceived with a stronger air burst and produced with more air release than the unaspirated plosives. By contrast, this hand burst gesture is not congruent with the phonetic features of aspirated affricates (i.e., a longer frication which is easily perceivable), since the quickly opened hand is not a good visual representation of the more gradual air release presenting a longer frication. Thus, while the hand burst gesture closely matches the auditory and articulatory properties of aspirated plosives, this is not the case for aspirated affricates.
The second goal of this PhD thesis is to assess whether the effects of audiovisual phonetic training with hand gestures will also be influenced by the degree of visual accessibility to phonetic features of the L2 phonemic contrasts, since as noted in section 1.2.3, the AV benefit on L2 perception depend upon the visual saliency of phonemes to be learned. Thus, two studies are proposed with Study 2 examining a barely visually accessible feature encoded by hand gestures and Study 3 dealing with a high visually accessible feature.

In Study 2, two pairs of sibilant consonants in Mandarin characterized as alveolar and retroflex (e.g., /ts/-/ʈʂ/, /s/-/ʂ/) will be trained. Acoustically, while some studies have reported different durational properties for the distinction of Mandarin alveolar-retroflex contrast (e.g., Feng, 1985; Shih & Ao, 1997), Svantesson (1983) suggested that the mean intensity of the fricative spectra could be an important perceptual cue to distinguish the contrast, with alveolar sounds having a higher mean intensity than their retroflex counterparts. Importantly, while the alveolar sounds are pronounced by putting the tip of the tongue at the teeth to form the constriction, when pronouncing the retroflex sounds, the tip of the tongue should be lifted to a position that is further back than alveolar sounds (Proctor et al., 2012). However, this articulatory information about the position of the tip of the tongue cannot be observed directly from the speakers’ face, which may cause difficulty for L2 learners to learn this feature. For example, Romance language speakers have been shown to mix up the alveolar and retroflex sounds when learning Mandarin (Chen et al., 2013). The subjects of this study are L1 Catalan speakers. In their L1 phonetic system, alveolar sounds such as /ts/ and /s/ exist; however, retroflex sounds do not. Thus, we assume that the learning of retroflex sounds will be difficult for Catalan speakers as well. We use a palm-down finger-lift gesture (see Figure 6) encoding the barely visually accessible feature.
accessible feature of lifting of the tip of the tongue for retroflex sounds to train the perception and production of these sounds.

**Figure 6.** A palm-down finger-lift gesture mimicking the lifting of the tip of the tongue for Mandarin retroflex sounds.

In Study 3, we will train L1 Mandarin speakers on L2 Catalan mid-closed and mid-open vowel contrasts (front unrounded /e/-/ɛ/ and back rounded /o/-/ɔ/). The visible articulatory feature of lip aperture is an important feature to distinguish these contrasts. However, Duanmu (2007) described that Mandarin has five vowel phonemes (/i/, /y/, /u/, /a/ and /a/) and some vowel sounds like /e/ and /o/ are considered as variants of mid vowel since they appear in certain phonetic environments (/e/ before /i/, /o/ in open syllables and after labials). Thus, we assume that for L1 Mandarin speakers, the Catalan vowel contrasts /e/-/ɛ/ and /o/-/ɔ/ will be difficult to learn. In order to train Mandarin speakers their on perception and production of these vowels, a set of four hand gestures was designed (see Figure 7) to mimic the articulatory features of the four target vowels considering the visible features of the lip aperture between the mid-closed and mid-open vowels and of the lip rounding between the two pairs of contrast.
Figure 7. Visuospatial hand gestures highlighting lip aperture and lip rounding. Upper panels show hand gestures for front unrounded vowel contrasts /e/-/ɛ/ (left and right panels) and the lower panels show hand gestures for back rounded vowel contrasts /o/-/ɔ/ (left and right panels).

2. Goals and Hypotheses

2.1 Goals of the dissertation

This PhD thesis has two main goals. The first goal is to investigate the learning of L2 sounds is affected by whether the gesture form matches or mismatches the target phonetic feature (Study 1). The second goal is to explore whether hand gestures highlighting different degrees of visually accessible articulation will influence the learning outcome (Studies 2 and 3)

All in all, the PhD thesis aims to explore the potential benefits of hand gestures encoding phonetic features in the learning of L2 phonemic contrasts at both the perception and production levels. A particular focus is placed on matching patterns between gesture form and phonetic features, as well as the role of visual accessibility of the target phonetic features encoded by hand gestures.

In order to achieve these goals, we designed three studies of audiovisual training with hand gestures, focusing on the two issues:

a. the role of the appropriateness of gesture form to target phonetic features (Study 1)
b. the role of the visual saliency of these target features encoded by hand gestures (Studies 2 and 3).

2.2 Hypotheses

Based on the literature review, we outline the main hypotheses of the present PhD thesis. The main overarching hypothesis of the three studies is that audiovisual training with hand gestures encoding phonetic features will facilitate the learning of L2 phonemic contrasts. And more specifically, we first hypothesize that hand forms which appropriately encode target phonetic features can boost phonetic training effects (see Study 1). Concerning the second research question, for audiovisual training with hand gestures encoding highly visually accessible articulation (Study 2), hand gestures could either enhance the visual input or distract the learner’s attention from the visually accessible articulation (i.e., the lip movements and lip and jaw aperture). While for training with hand gestures encoding barely visually accessible phonetic feature (Study 3), hand gestures will visually provide important information for articulation (e.g., tongue movement). Then, the second question will entertain the following two hypotheses: (a) hand gestures encoding highly visually accessible articulatory features might boost learning more than hand forms encoding barely accessible information; or by contrast, (b) hand gestures encoding barely visually accessible information might also yield beneficial training effects like those encoding highly visually accessible information.

2.2.1 Study 1

Research has shown that observing hand gestures mimicking pitch movements or rhythmic patterns can improve the learning of L2 suprasegmental features (e.g., Baills et al., 2019; Gluhareva & Prieto, 2017). However, less is known about the effects of hand gestures on the learning of novel phonemic contrasts.

Fifty Catalan native speakers undertook a short multimodal training session on two types of Mandarin Chinese consonants (plosives and affricates) in either of two conditions: Gesture and No Gesture. In the Gesture condition, a fist-to-open-hand gesture was used to mimic air burst, while the No Gesture condition included no such use of gestures. Crucially, while the hand gesture appropriately mimicked the air burst produced in plosives, this was not the case for affricates. Before and after training, participants were tested on two tasks, namely the identification task and the imitation task. Participants’ speech output was rated
by five Chinese native speakers. This study examines (a) whether hand gestures mimicking phonetic features can boost L2 segment learning by naïve learners, and (b) whether a mismatch between the hand gesture form and the target phonetic feature influences the learning effect. We expect that hand gestures will facilitate the learning only when they appropriately mimic the target phonetic features.

2.2.2 Study 2

Previous studies have shown that visual information from lip movements is an important source of information for speech perception (e.g., McGurk & MacDonald, 1976; Summerfield, 1992). Recent studies have shown that AV benefits on L2 perception depended upon the degree of visually accessible features to the target contrasts (e.g., Hazan et al., 2005). In this sense, certain L2 phonemic contrasts appear to be more difficult to learn than others due to the low level of visual accessibility. Since hand gestures are considered as a teaching tool in L2 classrooms to help L2 learners achieve target phonological features of pronunciation (e.g., Smotrova, 2017; Zhang, 2002), we ask whether adding hand gestures to audiovisual phonetic training to highlight a barely visually accessible information will compensate for the low level of visual accessibility of certain features to the phonemic contrasts.

The aim of this study is to investigate whether a palm-down finger-lift gesture mimicking the lifting of the tip of the tongue for retroflex sounds will facilitate the learning of Mandarin alveolar-retroflex contrast (/ts/-/ʈʂ/, /s/-/ʂ/). Importantly, the lifting of the tip of the tongue is not visually accessible to listeners, thus, the designed palm-down finger-lift gesture encodes an important but less visually accessible articulatory feature of the retroflex sounds.

In a between-subject experiment with a pre- and posttest design, 60 Catalan naïve learners without prior knowledge of Mandarin Chinese will watch audiovisual training materials under two conditions: No Gesture and Gesture. While participants in the Gesture condition will be exposed to target speech enriched with a palm-down finger-lift gesture mimicking the lifting of the tip of the tongue for Mandarin retroflex sounds, participants in the No Gesture condition will only see the lip movements and hear the speech. Before and after training, participants’ perception and production performance will be assessed using two tasks, namely an identification task and an imitation task. We expect that the Gesture group will outperform the No Gesture group on both perception and production.
2.2.2 Study 3

Although audiovisual perceptual trainings have been shown to have beneficial effects on the perception of L2 phonemic contrasts encoding high levels of visually accessible information (e.g., Hazan et al., 2005), it remains to be seen whether adding another layer of visual modality, namely hand gestures, will yield more improvements. The aim of Study 3 is to examine whether hand gestures encoding some highly visually accessible information, e.g., lip aperture and lip rounding, will boost the learning of two pairs of Catalan vowels contrasting in lip aperture (front unrounded /e/-/ɛ/ and back rounded /o/-/ɔ/).

In a between-subject experiment with a pre- and posttest design, 60 Chinese native speakers with no knowledge of Catalan will be trained in one of the two conditions: No Gesture and Gesture. In the Gesture condition, hand gestures mimicking different degrees of mouth opening and lip rounding are applied to corresponding sounds, while the No Gesture does not have this additional information. Before and after training, participants will be tested on three tasks, namely perceptual identification and word imitation. We have two hypotheses: (a) hand gestures will enhance the visual input and thus help the learning of the Catalan vowel contrasts; or (b) hand gestures will distract learner’s attention from the visually accessible articulation of the lip movements and lip and jaw aperture and then impair the learning.

Additionally, the training effect of hand gesture in this study will be compared with Study 2 to see whether the degree of visually accessible articulatory information encoded by hand gestures will yield different learning outcomes. We have two hypotheses: (a) training with hand gestures encoding highly visually accessible articulatory features will help the learning of L2 speech sounds more than training with barely visually accessible information; or by contrast, (b) training with hand gestures encoding barely visually accessible information yield more beneficial training effects.
3. Experimental studies

3.1 Study 1: Training with a fist-to-open hand gesture on the learning of novel aspiration contrasts

3.1.1 Research questions

The goal of Study 1 is twofold: first, to assess whether using a hand burst gesture which visually depicts the air burst and aspiration will facilitate the learning of Mandarin aspiration contrasts (plosives and affricates); and second, to test whether the presence of a mismatch between the hand gesture and the target phonetic features influences phonological learning. Crucially, while the hand burst gesture used in the study closely matches the properties of aspirated plosives, this same gesture is a poor match for the properties of aspirated affricates.

We thus address the following two research questions: Does audiovisual training with hand gestures encoding aspiration improve the perception and production of Mandarin aspiration contrasts? Further, does the appropriateness of the hand gesture for the phonetic features in question have an effect on the learning outcome?

3.1.2 Participants

Fifty Catalan-dominant speakers (N = 50, 42 females and eight males; M(age) = 20.90 years, SD = 2.496, SE = 0.250) were recruited from a public university in Barcelona. Prior to the experiment, participants were asked to answer a questionnaire about their age, gender, and linguistic and musical background (see Appendix). All the participants reported more than 75% use of Catalan in daily verbal communication, and none of them had studied Mandarin Chinese before. They were then randomly assigned to one of the two training conditions, namely, NG (n = 25, 21 females and four males) and G (n = 25, 21 females and four males). They signed a written consent giving permission to process their data and received a small amount of money in compensation for their participation.

3.1.3 Materials

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

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

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

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

Before the recordings, the two instructors were trained to use the same hand burst gesture to accompany their production of aspirated consonants embedded in words: This
involved simultaneously raising their two hands in clenched fists to a position somewhat higher than their shoulders and then opening them suddenly to splay their fingers with a forward movement toward the camera in order to illustrate the air burst (see Figure 5). Then, they immediately put back their hands in the rest position as they finished pronouncing the word.

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

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

The second block was intended to train the words contrastively in pairs. First, the Catalan-adapted orthography of the minimal pair was presented on the screen. Then, each minimal pair was produced first by one instructor and then by the other. This sequence was followed by a black screen. The second block was repeated twice with different trial orders. Therefore, in total, each minimal pair was presented 4 times (1 word pair $\times$ 2 instructors $\times$ 2 times).
Auditory stimuli for the pre- and posttest identification task. For the identification task used in the pre- and post-test, six pairs of Mandarin disyllabic words featuring the aspiration contrast in word-initial position were chosen (see Appendix 3). Half of the words were included in the training session video, and the other half were not.

The audio recordings were performed in the radio studio using professional equipment and later edited with Audacity 2.1.2 software. All words were recorded twice at a normal speech rate by the same two instructors. Later, the clearest and most natural-sounding samples were selected for the final 24 audio files (12 words × 2 instructors). For the 12 words in both pre- and posttests, half of the audio recordings were spoken by one instructor and the other half were spoken by the other instructor, with an obligatory alternation between pretest and posttest. That is, if in the pretest, the audio recording of a word was produced by one instructor, the recording of the same word for the post-test should be produced by the other instructor. The audio files were then uploaded to the online survey platform SurveyGizmo with the order of items automatically randomized within each test by the software.

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

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

3.1.4 Experimental procedure

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

A summary of the experimental procedure can be seen in Figure 3. Prior to the experiment, participants signed a consent form and answered a questionnaire about their age, gender, and linguistic and musical background, as noted above. They were then randomly assigned to one of the two conditions, NG or G. The experiment started with each participant watching the familiarization video (3 min 11 s). Next, they performed the
identification and imitation tasks making up the pretest. Immediately after the pretest, they watched the short training video (5 min 36 s). Finally, they completed the posttest, which consisted of the same tasks performed in the pretest. Altogether, the experiment lasted about 25 min.

Figure 8. Experimental procedure.

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

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

**Training session.** After the pretest, participants watched the training video involving six pairs of words featuring unaspirated and aspirated consonants. In the NG condition, participants watched the instructors utter the words, whereas in the G condition, participants watched the instructors performing the gesture as they simultaneously produced the aspirated consonants. In both conditions, the participant was expected to remain motionless and silent as they watched the training video.

### 3.1.5 Coding of data

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

**Identification task.** Participants’ responses were assessed according to a binary rating system whereby a correct answer was given a score of ‘1’ and an incorrect answer, ‘0’.

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

All the recordings were presented to the raters randomly so that they were unaware of the training conditions and tests. Before rating, all raters were trained in a 30-min session with some examples so as to become familiar with the evaluation system. Raters first rated the general pronunciation of all audio recordings and then listened to the audio samples again and rated the consonant feature accuracy. For general pronunciation accuracy, raters listened to each word and evaluated the pronunciation of the word on a 9-point Likert scale from 9 = definitely accurate to 1 = not accurate at all by focusing on the production of both consonants and vowels as well as the lexical tones. For the accuracy of the consonantal feature, they were asked to pay attention to the word-initial segment and likewise evaluate it on a 9-point Likert scale. Raters were asked to focus on the strength of air burst for assessing the accuracy of aspiration features for plosives and on the duration of the airflow for assessing the accuracy of frication features for affricates. For instance, for /pʰ/, raters should give a high score (7–9) if they heard a /pʰ/-like sound with a strong air burst, a middle score (4–6) if they could not recognize whether the sound was /p/ or /pʰ/, and a low score (1–3) for a /p/-like sound with a weak air burst. Additionally, raters assessed the accuracy of consonantal feature regardless of whether the consonant produced was the target consonant or not. For example, if the target sound was /tʰa/ but the participant produced /kʰa/, the rater still gave the output an “accurate” rating because the aspiration feature was accurately pronounced even if the consonant itself was not.

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

**Musical background.** As musical experience may play a role in phonological learning (see Chobert & Besson, 2013, for a review), we controlled for this factor to ensure that there was no difference in this regard between the two training groups. First, adapting Boll-Avetisyan, Bhatara and Höhle (2017) procedure, a musical expertise score was obtained for each participant by coding their answers to the musical background questionnaire as follows: (a) for the years spent studying music, the score equaled the number of years reported; (b) for the number of instruments played, 1 point was given for each instrument reported; (c) regarding the amount of time they reported spending on a regular basis singing or listening to music, 5 points were given if the participant reported daily frequency, 4 points for 5-6 days per week, 3 for 3-4 days per week, 2 for 1-2 days per week, 1 for ‘occasionally’, and 0 for ‘never’. The sum of all these scores constituted a ‘Musical Expertise’ score. In addition, the musical skill of participants was rated by means of a self-assessment tool (following Law & Zentner, 2012) that yielded a score whereby 1 = nonmusician, 2 = music-loving nonmusician, 3 = amateur musician, 4 = semiprofessional musician, and 5 = professional musician. This constituted their “self-perceived musical skills” score.

### 3.1.6 Statistical analyses

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

A generalized linear mixed model (GLMM) was applied to the following outcome measures using the `glmmTMB` package (Brooks et al., 2017) in R: (a) identification accuracy, that is, the participant’s score for each item in the identification task; (b) pronunciation accuracy, that is, the participant’s general pronunciation scores for each item rated by each rater in the imitation task; or (c) consonantal feature accuracy, that is, the participant’s consonantal feature scores for each item rated by each rater in the imitation task. The fixed factors were condition (two levels: NG and G), test (two levels: pre- and
posttest), and consonant type (two levels: plosives and affricates), as well as their interactions. A series of GLMMs using different random effects structures were modeled, from the most complex random effects structure to a marginal model with no random effects. All the structures that did not produce any converge problems were compared using the function *compare performance* from the *performance* package (Lüdecke et al., 2019) to identify the model that best fitted our data. For identification accuracy, the best-fitting model was the one including a random intercept both for participant and for item (i.e., (1 | Participant) + (1 | Item)). In the case of the pronunciation accuracy, the best-fitting model was the one with the random effects structure including random slopes for Test, Consonant Type and Test × Consonant Type by Participant, for Condition by Item, and for Consonant Type by Rater (i.e., (1 + Test * Consonant Type | Participant) + (1 + Condition | Item) + (1 + Consonant Type | Rater)). In the case of the consonantal feature accuracy, the best-fitting model was the one including a random slope for Test, Consonant Type and Test × Consonant Type by Participant, a random slope for both Condition and Test by Item, a random intercept for rater (i.e., (1 + Test * Consonant Type | Participant) + (1 + Condition + Test | Item) + (1 | Rater)). In the results below, the omnibus test results are provided, plus the output of a series of Bonferroni pairwise tests performed with the *emmeans* package (Lenth et al., 2019), which include a measure of effect size by using Cohen’s *d*.

### 3.1.7 Results

The results of the three GLMMs with the three outcome measures (Identification Accuracy, Pronunciation Accuracy and Consonantal Feature Accuracy) as the dependent variables are illustrated in Table 1.

**Identification accuracy.** Figure 9 shows the mean Identification Accuracy rates across Condition (NG and G), Test (Pretest and Posttest) and Consonant Type (Plosives and Affricates). Results of the GLMM with the Identification Accuracy as the dependent variable (see Table 1) revealed a significant main effect of Consonant Type (χ²(1) = 15.557, \(p < .001\)) and Test (χ² (1) = 4.842, \(p = .028\)), indicating that participants’ performance differed significantly from pretest to posttest and varied significantly across the two types of consonant. No main effect of Condition was found (χ² (1) = 1.117, \(p = .291\)). Post-hoc comparisons revealed that participants improved significantly after training in general (\(d = 0.13, p = .028\)) and identified plosives significantly better than affricates (\(d = 0.47, p < .001\)). No significant two-way interactions between Condition × Test (χ² (1) = 0.697, \(p < .
The three-way interaction between Condition × Test × Consonant Type ($\chi^2(1) = 2.078, p = .149$) were found, which suggests that (a) training with gestures did not yield more of an improvement for the perception of nonnative aspiration contrast than training without gestures and (b) the appropriateness of the gesture (i.e., whether or not the form or manner of the hand gesture seemed to visually mimic the phonetic properties of the target phonemes) did not influence the training effect.

![Figure 9](image)

**Figure 9.** Mean Identification Accuracy rates across Condition (NG and G), Test (Pretest and Posttest) and Consonant Type (Plosives and Affricates). Error bars indicate $\pm 2 \text{SE}$. 

**Pronunciation accuracy.** The Figure 10 shows the mean Pronunciation Accuracy obtained in the imitation task across Condition (NG and G), Test (Pretest and Posttest) and Consonant Type (Plosives and Affricates). Results of the GLMM with Pronunciation Accuracy as the dependent variable (see Table 1) revealed a main effect of Consonant Type ($\chi^2(1) = 7.575, p = .006$), showing that participants’ general pronunciation was significantly different depending on whether they pronounced plosives or affricates. Participants showed better performance on the plosives than on the affricates ($d = -.39, p = .004$). Two significant two-way interactions were found: Condition × Consonant Type ($\chi^2(1) = 7.963, p = .005$), indicating that the performance of G and NG groups differed significantly depending on the consonant type; and Test × Consonant Type ($\chi^2(1) = 6.106, p < .013$), indicating that participants’ performance on different types of consonants differed...
significantly from pre- to posttest. Crucially, a significant three-way interaction was found in Condition × Test × Consonant Type ($\chi^2(1) = 5.855, p = .016$), revealing that the participants in both G and NG groups were significantly different from pretest to posttest depending on the type of consonant. Post-hoc comparisons (see Table 2) showed that participants’ general pronunciation of the words containing plosives improved significantly in the G group ($d = 0.27, p = .023$) but not in the NG group, while for words containing affricates, performance did not improve in either NG or G group. These results showed that training with gestures that appropriately mimicked phonetic features significantly helped the general pronunciation of the target words compared to training without gestures and training with gestures that inappropriately represented phonetic behaviors. These results suggest that when gestures appropriately represent the auditory and articulatory properties of the target phonemes, participants can improve their pronunciation performance. By contrast, if a gesture does not visually match the auditory and articulatory properties of the target phonemes, this may not help the learning.

![Graph showing pronunciation accuracy](image)

**Figure 10.** Mean Pronunciation Accuracy across Condition (NG and G), Test (Pretest and Posttest) and Consonant Type (Plosives and Affricates). Error bars indicate ±2 SEs.

**Consonantal feature accuracy.** The Figure 11 shows mean Consonantal Feature Accuracy across Condition (NG and G), Test (Pretest and Posttest) and Consonant Type (Plosives and Affricates). Results of a GLMM with Consonantal Feature Accuracy as the
dependent variable (see Table 1) revealed a two-way interaction between Test × Consonant Type ($\chi^2(1) = 25.758, p < .001$), suggesting that participants’ performance differed significantly from pretest to posttest according to the type of consonantal feature. A significant three-way interaction between Condition × Test × Consonant Type ($\chi^2(1) = 11.279, p = .001$) was obtained. The post-hoc results (see more details in Table 2) showed that the improvement from pre to posttest for plosives was present only in the G condition ($d = 0.33, p < .001$), but not for any combination including NG or affricates data, suggesting that participants’ performance of consonantal feature improved only when gestures matched the properties of the target phonemes. When participants were trained with matching gestures, their performance on consonantal feature improved significantly from pretest to posttest; however, when they were trained with mismatching gestures or without any gestures, their performance did not improve after training. Taken together, these results strongly suggest that training with gestures had asymmetric effects depending on whether the gestures appropriately mimicked the phonetic features of the target phonemes or not.

Figure 11. Mean Consonantal Feature Accuracy across Condition (NG and G), Test (Pretest and Posttest) and Consonant Type (Plosives and Affricates). Error bars indicate 2 SEs.

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

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

3.2 Study 2: Training with a palm-down finger-lift gesture for the learning of alveolar-retroflex contrasts

3.2.1 Research questions

Study 2 examines the effects of a palm-down finger-lift gesture which encodes the lifting of the tip of the tongue for retroflex sounds on the learning of L2 Mandarin alveolar-retroflex contrasts (/ts/-/ʈʂ/, /s/-/ʂ/). We will expand the work performed by previous studies by assessing not only the role of hand gestures in L2 perception but also in production and explore whether adding hand gestures encoding phonetic features to the audiovisual training will compensate for the low level of visual accessibility of certain features to the phonemic contrasts.

We thus address the following research question: Does audiovisual training with hand gestures highlighting barely visually accessible articulatory information, the lifting of the tip of the tongue, facilitates the learning of L2 alveolar-retroflex contrasts by naïve Catalan learners?
3.2.2 Participants

In a between-subject experiment with a pre- and posttest design, 60 Catalan native speakers will be recruited. Prior to the experiment, participants will be asked to answer a questionnaire about their age, gender, and linguistic and musical background (see Appendix 2). All the participants should report their daily Catalan use above 75% and that they have never learned Mandarin before. Then they will be randomly assigned to one of the two training conditions: (a) the No Gesture (NG) condition, in which participants will watch the instructors producing training stimuli without any body movements; and (b) the Gesture (G) condition, in which participants will watch the instructors performing hand gestures for the target sounds. In both training conditions, participants will be asked to remain still and silent when watching the training videos.

3.2.3 Materials

The experimental materials (see Appendix 4) for Study 2 will be similar to those in Study 1 with the only difference being that the minimal pairs of disyllabic words will contain consonantal contrasts involving alveolar-retroflex sounds in word-initial position (e.g., siùmiáo ‘sketch’ vs. shùmiáo ‘sapling’). Similarly, two training sets of videos will be produced for the two training conditions: (a) for the NG condition, the instructors will produce the training stimuli without any hand gestures; and (b) for the G condition, the instructors will perform the hand gestures along with the words containing retroflex sounds (see Figure 6), but they do not gesture for the word containing alveolar sounds.

3.2.4 Experimental procedure

The full experimental procedure will be carried out in a quiet room which is the same as described in Study 1. A summary is shown in Figure 12.

Figure 12. Experimental procedure.
3.2.5 Coding of Data

After the experiment, participants’ responses from the identification task will be assessed according to a binary rating system whereby a correct answer was given a score of ‘1’ and an incorrect answer, ‘0’. The recordings obtained in the imitation task will be rated by five Mandarin native speakers, who will be blind to the experimental conditions. They will be asked to assess two pronunciation variables from each audio file using a 9-point Likert Scale: (a) the general pronunciation accuracy of the target word; and (b) the accuracy of the target consonants. In addition, the “Musical Expertise” and “Self-Perceived Musical Skills” scores will be obtained by coding the questionnaire as described in section 3.1.5.

3.3 Study 3: Training with hand gestures encoding lip aperture and lip rounding for the learning of close-mid and open-mid vowel contrasts

3.3.1 Research questions

This study will assess whether hand gestures mimicking a visible articulatory feature (i.e., lip aperture and lip rounding) will yield better learning outcomes for L2 Catalan close-mid and open-mid vowel contrasts (front unrounded /e/-/ɛ/ contrast and back rounded /o/-/ɔ/ contrast).

We thus address the following research question: Does audiovisual training with visuospatial hand gestures encoding highly visually accessible information, the lip aperture and lip rounding, facilitates the learning of two pairs of L2 novel vowels contrasting in lip aperture by naïve Mandarin learners?

3.3.2 Participants

In a between-subject experiment with a pre- and posttest design, 60 Mandarin native speakers without prior knowledge of Catalan will be recruited. All of the participants will answer the same questionnaire as Study 1, which asks for participants’ personal information such as age, gender, linguistic experience and musical background (see Appendix 2). Then they will be randomly assigned to one of the two training groups: (a) in the No Gesture (NG) group, participants will watch the instructors producing training stimuli without any body movements; and (b) in the Gesture (G) group, participants will
watch the instructors performing hand gestures for the target sounds.

### 3.3.3 Materials

The experiment consists of a familiarization phase introducing the phonetic properties of the Catalan closed-mid and open-mid vowel contrasts. Then, two pretest tasks (identification and imitation tasks) will be followed by a short training session. Next, the posttest consists of the same identification and imitation tasks performed in the pretest. The types of materials used in aforementioned phases will be similar to those used in Study 1, with the only difference being that the vowel minimal pairs of Catalan words contrast in lip aperture (e.g., Déu ‘god’ vs. deu ‘ten’) (see Appendix 5).

Two training sets of videos will be produced for the two conditions: (a) for NG condition, the instructors produce the training stimuli without any hand gestures, whereas for G condition, the instructors will perform the hand gestures so that the finger positions mimic the aperture and rounding of the lips of the target vowels (see Figure 7). As noted in the figure, the distance in space between thumb and other fingers depicts the degree of mouth aperture for the two pairs of contrasts.

### 3.3.4 Experimental procedure

The full experimental procedure will be carried out in a quiet room which is the same as described in Study 1. A summary is shown in Figure 13.

![Figure 13. Experimental procedure.](image)

### 3.3.5 Coding of data

Participants’ responses in the identification task will be assessed according to a binary rating system whereby a correct answer will be given a score of ‘1’ and an incorrect ‘0’.

The recordings in the imitation task will be assessed by 5 Catalan native speakers using a 9-point Likert scale by focusing on the general pronunciation accuracy of the target
words and the target vowels. Moreover, phoneme accuracy will be measured by means of acoustic analyses (namely, formant analyses of F1, F2 and F3).

The “Musical Expertise” and “Self-Perceived Musical Skills” scores will be obtained for each participant by coding the questionnaire as described in section 3.1.5.

All in all, we hope this PhD Thesis will provide a more comprehensive role of hand gestures in L2 segment learning. From a pedagogical perspective, we hope this work can encourage L2 teachers to use more hand gestures in the classroom when teaching difficult sounds, and also help them to be aware of choosing or designing appropriate hand gestures considering the properties of the target sounds. Moreover, since hand gestures are widely used in the clinical field to help people with apraxia to better pronounce speech sounds, we hope our work will provide evidence of the interaction between speech sounds and hand gestures, and thus support the incorporation of hand gestures in therapy for apraxia.

4. Work plan

<table>
<thead>
<tr>
<th>Time Period</th>
<th>Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>May 2020 – Aug. 2020</td>
<td>Oral defense of Research Plan</td>
</tr>
<tr>
<td></td>
<td>Research design of Study 2</td>
</tr>
<tr>
<td></td>
<td>Creation of materials of Study 2</td>
</tr>
<tr>
<td></td>
<td>Data analysis of Study 2</td>
</tr>
<tr>
<td></td>
<td>Oral presentation in GESPIN (September 7-9, Stockholm)</td>
</tr>
<tr>
<td>Jan. 2021 – Apr. 2021</td>
<td>Writing the scientific article based on Study 2</td>
</tr>
<tr>
<td></td>
<td>Research design of Study 3</td>
</tr>
<tr>
<td></td>
<td>Creation of materials of Study 3</td>
</tr>
<tr>
<td>May 2021 – Aug. 2021</td>
<td>Submission of scientific article based on Study 2</td>
</tr>
<tr>
<td></td>
<td>Conduction of experiment for Study 3</td>
</tr>
<tr>
<td></td>
<td>Data analysis of Study 3</td>
</tr>
<tr>
<td></td>
<td>Oral presentation of Study 1 in the 30th EuroSLA in Barcelona (accepted)</td>
</tr>
<tr>
<td></td>
<td>Writing the scientific article based on Study 3</td>
</tr>
<tr>
<td></td>
<td>Oral presentation of Study 2 in one or two conferences</td>
</tr>
</tbody>
</table>
5. References


ntonation more than length in auditory judgments of non-native phonemic contrasts. *Collabra: Psychology, 3*(1), 7. https://doi.org/10.1525/collabra.76


## Appendix

### 1 Summary table of hand gestures cueing phonemes

<table>
<thead>
<tr>
<th>Articulatory feature</th>
<th>Segment</th>
<th>Gesture shape</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mouth shape - open mouth wide</td>
<td>/æ/</td>
<td></td>
<td>Hudson, 2011</td>
</tr>
<tr>
<td>Mouth shape - lip roundness</td>
<td>/u/</td>
<td></td>
<td>Hudson, 2011</td>
</tr>
<tr>
<td>Teeth, and tongue</td>
<td>Mandarin alveolar</td>
<td></td>
<td>Zhang, 2002</td>
</tr>
<tr>
<td>Tongue</td>
<td>Mandarin retroflex</td>
<td></td>
<td>Zhang, 2002</td>
</tr>
<tr>
<td>Articulatory Feature</td>
<td>Language</td>
<td>Region</td>
<td>Author</td>
</tr>
<tr>
<td>----------------------</td>
<td>-----------------</td>
<td>--------------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>Tongue</td>
<td>Mandarin</td>
<td>alveolar-palatal</td>
<td>Zhang, 2002</td>
</tr>
<tr>
<td>Tongue</td>
<td>Mandarin</td>
<td>alveolar nasal</td>
<td>Zhang, 2002</td>
</tr>
<tr>
<td>Tongue</td>
<td>Mandarin</td>
<td>velar nasal</td>
<td>Zhang, 2002</td>
</tr>
<tr>
<td>Articulatory feature</td>
<td>English</td>
<td>unreleased stops</td>
<td>Armand &amp; Touhami, 2016</td>
</tr>
<tr>
<td>Articulatory feature</td>
<td>English</td>
<td>released stops</td>
<td>Armand &amp; Touhami, 2016</td>
</tr>
<tr>
<td>Mouth Shape - Lip Roundness</td>
<td>Spanish /u/</td>
<td>Hoetjes, van Maastricht &amp; van der Heijden, 2019</td>
<td></td>
</tr>
<tr>
<td>-----------------------------</td>
<td>-------------</td>
<td>--------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Mouth Shape - Lip Roundness</td>
<td>Spanish /θ/</td>
<td>Hoetjes, van Maastricht &amp; van der Heijden, 2019</td>
<td></td>
</tr>
<tr>
<td>Tongue</td>
<td>English /r/</td>
<td>Lan &amp; Wu, 2013</td>
<td></td>
</tr>
<tr>
<td>Tongue</td>
<td>English /l/</td>
<td>Lan &amp; Wu, 2013</td>
<td></td>
</tr>
<tr>
<td>Mouth Shape - Lip Openness (Small)</td>
<td>Vowels with Small Lip Openness</td>
<td>Rusiewicz &amp; Rivera, 2017</td>
<td></td>
</tr>
<tr>
<td>----------------------------------</td>
<td>-------------------------------</td>
<td>-------------------------</td>
<td></td>
</tr>
<tr>
<td>Mouth Shape - Lip Openness (Large)</td>
<td>Vowels with Large Lip Openness</td>
<td>Rusiewicz &amp; Rivera, 2017</td>
<td></td>
</tr>
<tr>
<td>Tongue</td>
<td>/r/</td>
<td>Rusiewicz &amp; Rivera, 2017</td>
<td></td>
</tr>
</tbody>
</table>

2 Linguistic and musical background questionnaire

Linguistic experience
1. What percentage of CATALAN do you use in your daily life?
2. Apart from CATALAN and SPANISH, which language(s) do you speak?
3. Have you ever studied Japanese?

Musical experience
1. How many years of musical education have you ever received?
2. Do you play any instruments? If your answer is yes, answer. If not, move on to 4.
3. Which instrument(s) do you play?
4. How often do you sing or listen to music?
   A. Every day   B. 5-6 days per week   C. 3-4 days per week
   D. 1-2 days per week   E. Occasionally   F. Never
5. Which one of the following best describes you?
   A. I’m a non-musician
   B. I’m a music-loving non-musician
   C. I’m an amateur musician
   D. I’m a semi-professional musician
   E. I’m a professional musician

3 Experimental materials of Study 1

3.1 Stimuli for the training session

<table>
<thead>
<tr>
<th>Word</th>
<th>Target</th>
<th>Consonant Type</th>
<th>Catalan adaptation</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>pí yán</td>
<td>pʰ</td>
<td>plosive</td>
<td>pʰi yan</td>
<td>dermatitis</td>
</tr>
<tr>
<td>bì yán</td>
<td>p</td>
<td>plosive</td>
<td>pi yan</td>
<td>rhinitis</td>
</tr>
<tr>
<td>tān liàn</td>
<td>tʰ</td>
<td>plosive</td>
<td>tʰan lian</td>
<td>greed</td>
</tr>
<tr>
<td>dān liàn</td>
<td>t</td>
<td>plosive</td>
<td>tan lian</td>
<td>one-side love</td>
</tr>
<tr>
<td>kǒu liáng</td>
<td>kʰ</td>
<td>plosive</td>
<td>kʰou liang</td>
<td>ration</td>
</tr>
<tr>
<td>gǒu liáng</td>
<td>k</td>
<td>plosive</td>
<td>kou liang</td>
<td>dog food</td>
</tr>
<tr>
<td>cuò wù</td>
<td>tʰ</td>
<td>affricate</td>
<td>tsʰuo u</td>
<td>error</td>
</tr>
<tr>
<td>zuò wù</td>
<td>ts</td>
<td>affricate</td>
<td>tsuo u</td>
<td>crop</td>
</tr>
<tr>
<td>qīng lǐ</td>
<td>tʰ</td>
<td>affricate</td>
<td>tsjʰing li</td>
<td>to clean</td>
</tr>
<tr>
<td>jīng lǐ</td>
<td>tɕʰ</td>
<td>affricate</td>
<td>tsj引擎 li</td>
<td>manager</td>
</tr>
<tr>
<td>chū lǐ</td>
<td>tɕʰ</td>
<td>affricate</td>
<td>txʰu li</td>
<td>to stand</td>
</tr>
<tr>
<td>zhū lǐ</td>
<td>tɕʰ</td>
<td>affricate</td>
<td>txu li</td>
<td>boosting</td>
</tr>
</tbody>
</table>

Note. In the “Word” column, all items are written in pinyin (standard Romanized Chinese).
### 3.2 Stimuli for the identification task

<table>
<thead>
<tr>
<th>Word</th>
<th>Target</th>
<th>Trained</th>
<th>Catalan adaptation</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>pí yán</td>
<td>pʰ</td>
<td>Yes</td>
<td>pʰi yan</td>
<td>dermatitis</td>
</tr>
<tr>
<td>bì yán</td>
<td>p</td>
<td>Yes</td>
<td>pi yan</td>
<td>rhinitis</td>
</tr>
<tr>
<td>tān liàn</td>
<td>tʰ</td>
<td>Yes</td>
<td>tʰan lian</td>
<td>greed</td>
</tr>
<tr>
<td>dān liàn</td>
<td>t</td>
<td>Yes</td>
<td>tan lian</td>
<td>one-side love</td>
</tr>
<tr>
<td>kǔ li</td>
<td>kʰ</td>
<td>No</td>
<td>kʰu li</td>
<td>labor</td>
</tr>
<tr>
<td>gǔ li</td>
<td>k</td>
<td>No</td>
<td>ku li</td>
<td>encouragement</td>
</tr>
<tr>
<td>cuò wù</td>
<td>tsʰ</td>
<td>Yes</td>
<td>tsʰuo u</td>
<td>error</td>
</tr>
<tr>
<td>zuò wù</td>
<td>ts</td>
<td>Yes</td>
<td>tsuo u</td>
<td>crop</td>
</tr>
<tr>
<td>qiè hé</td>
<td>teʰ</td>
<td>No</td>
<td>tsjʰe he</td>
<td>stuffed eggplant</td>
</tr>
<tr>
<td>jiè hé</td>
<td>te</td>
<td>No</td>
<td>tsje he</td>
<td>tuberculosis</td>
</tr>
<tr>
<td>chǎn shì</td>
<td>ʈʰ</td>
<td>No</td>
<td>txʰan xi</td>
<td>to elucidate</td>
</tr>
<tr>
<td>zhǎn shì</td>
<td>ʈ</td>
<td>No</td>
<td>txan xi</td>
<td>to demonstrate</td>
</tr>
</tbody>
</table>

### 3.3 Stimuli for the imitation task

<table>
<thead>
<tr>
<th>Word</th>
<th>Target</th>
<th>Trained</th>
<th>Meaning</th>
<th>Words</th>
<th>Target</th>
<th>Trained</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>pá shǒu</td>
<td>pʰ</td>
<td>No</td>
<td>thief</td>
<td>cōng yóu</td>
<td>tsʰ</td>
<td>No</td>
<td>scallion oil</td>
</tr>
<tr>
<td>bà shǒu</td>
<td>p</td>
<td>No</td>
<td>handle</td>
<td>zōng yóu</td>
<td>ts</td>
<td>No</td>
<td>palm oil</td>
</tr>
<tr>
<td>tào lù</td>
<td>tʰ</td>
<td>No</td>
<td>strategy</td>
<td>qǐng lǐ</td>
<td>teʰ</td>
<td>Yes</td>
<td>to clean</td>
</tr>
<tr>
<td>dào lù</td>
<td>t</td>
<td>No</td>
<td>road</td>
<td>jǐng lǐ</td>
<td>te</td>
<td>Yes</td>
<td>manager</td>
</tr>
<tr>
<td>kǒu liáng</td>
<td>kʰ</td>
<td>Yes</td>
<td>ration</td>
<td>chǔ lǐ</td>
<td>ʈʰ</td>
<td>Yes</td>
<td>to stand</td>
</tr>
<tr>
<td>gǒu liáng</td>
<td>k</td>
<td>Yes</td>
<td>dog food</td>
<td>zhǔ lǐ</td>
<td>ʈ</td>
<td>Yes</td>
<td>boosting</td>
</tr>
</tbody>
</table>

### 4 Experimental materials of Study 2

#### 4.1 Stimuli for the training session

<table>
<thead>
<tr>
<th>Word</th>
<th>Target</th>
<th>Meaning</th>
<th>Word</th>
<th>Target</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>zhōng yì</td>
<td>ʈʰ</td>
<td>favorite</td>
<td>shī wén</td>
<td>ʂ</td>
<td>poem</td>
</tr>
</tbody>
</table>
### 4.2 Stimuli for the identification task

<table>
<thead>
<tr>
<th>Word</th>
<th>Trained</th>
<th>Meaning</th>
<th>Word</th>
<th>Trained</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>zhōng yì</td>
<td>Yes</td>
<td>favorite</td>
<td>shāng yè</td>
<td>No</td>
<td>business</td>
</tr>
<tr>
<td>zhōng yì</td>
<td>Yes</td>
<td>variety show</td>
<td>sāng yè</td>
<td>No</td>
<td>folium mori</td>
</tr>
<tr>
<td>zhú lián</td>
<td>Yes</td>
<td>bamboo curtain</td>
<td>shuǐ hé</td>
<td>No</td>
<td>who</td>
</tr>
<tr>
<td>zú lián</td>
<td>Yes</td>
<td>football association</td>
<td>sū yǔ</td>
<td>S</td>
<td>saying</td>
</tr>
<tr>
<td>zhǎo dào</td>
<td>No</td>
<td>find</td>
<td>shì zhě</td>
<td>Yes</td>
<td>envoy</td>
</tr>
<tr>
<td>zǎo dào</td>
<td>No</td>
<td>early rice</td>
<td>sì zhě</td>
<td>Yes</td>
<td>dead</td>
</tr>
<tr>
<td>zhǐ fù</td>
<td>No</td>
<td>uniform</td>
<td>shù dí</td>
<td>Yes</td>
<td>clarinet</td>
</tr>
<tr>
<td>zì fù</td>
<td>No</td>
<td>character</td>
<td>sù dí</td>
<td>Yes</td>
<td>old enemy</td>
</tr>
</tbody>
</table>

### 4.3 Stimuli for the imitation task

<table>
<thead>
<tr>
<th>Word</th>
<th>Trained</th>
<th>Meaning</th>
<th>Word</th>
<th>Trained</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>zhuān yán</td>
<td>No</td>
<td>specialize in</td>
<td>shī wén</td>
<td>Yes</td>
<td>poem</td>
</tr>
<tr>
<td>zuān yán</td>
<td>No</td>
<td>dig into</td>
<td>sī wén</td>
<td>Yes</td>
<td>gentle</td>
</tr>
<tr>
<td>zhá guō</td>
<td>No</td>
<td>fryer</td>
<td>shú yǔ</td>
<td>Yes</td>
<td>idiom</td>
</tr>
<tr>
<td>zá guō</td>
<td>No</td>
<td>fail</td>
<td>sū yǔ</td>
<td>Yes</td>
<td>saying</td>
</tr>
<tr>
<td>zhǐ téng</td>
<td>Yes</td>
<td>stop pain</td>
<td>shǎn guāng</td>
<td>No</td>
<td>glitter</td>
</tr>
<tr>
<td>zǐ téng</td>
<td>Yes</td>
<td>Chinese wisteria</td>
<td>sān guāng</td>
<td>No</td>
<td>astigmatism</td>
</tr>
<tr>
<td>zhào piàn</td>
<td>Yes</td>
<td>photo</td>
<td>shù miáo</td>
<td>No</td>
<td>sapling</td>
</tr>
<tr>
<td>zào piàn</td>
<td>Yes</td>
<td>soap flake</td>
<td>sù miáo</td>
<td>No</td>
<td>sketch</td>
</tr>
</tbody>
</table>
5 Experimental materials of Study 3

5.1 Stimuli for the training session

<table>
<thead>
<tr>
<th>Word</th>
<th>Target</th>
<th>Vowel Type</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>bé</td>
<td>e</td>
<td>mid-close</td>
<td>well</td>
</tr>
<tr>
<td>be</td>
<td>ɛ</td>
<td>mid-open</td>
<td>letter b</td>
</tr>
<tr>
<td>més</td>
<td>e</td>
<td>mid-close</td>
<td>more</td>
</tr>
<tr>
<td>mes</td>
<td>ɛ</td>
<td>mid-open</td>
<td>month</td>
</tr>
<tr>
<td>sol</td>
<td>o</td>
<td>mid-close</td>
<td>sun</td>
</tr>
<tr>
<td>sòl</td>
<td>ɔ</td>
<td>mid-open</td>
<td>ground</td>
</tr>
<tr>
<td>fon</td>
<td>o</td>
<td>mid-close</td>
<td>phoneme</td>
</tr>
<tr>
<td>font</td>
<td>ɔ</td>
<td>mid-open</td>
<td>source</td>
</tr>
</tbody>
</table>

5.2 Stimuli for the identification task

<table>
<thead>
<tr>
<th>Word</th>
<th>Trained</th>
<th>Meaning</th>
<th>Word</th>
<th>Trained</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>déu</td>
<td>e</td>
<td>god</td>
<td>món</td>
<td>ɔ</td>
<td>world</td>
</tr>
<tr>
<td>deu</td>
<td>ɛ</td>
<td>ten</td>
<td>mon</td>
<td>o</td>
<td>my</td>
</tr>
<tr>
<td>mès</td>
<td>e</td>
<td>more</td>
<td>sol</td>
<td>o</td>
<td>sun</td>
</tr>
<tr>
<td>mes</td>
<td>ɛ</td>
<td>month</td>
<td>sòl</td>
<td>ɔ</td>
<td>ground</td>
</tr>
</tbody>
</table>

5.3 Stimuli for the imitation task

<table>
<thead>
<tr>
<th>Word</th>
<th>Trained</th>
<th>Meaning</th>
<th>Word</th>
<th>Trained</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>bé</td>
<td>e</td>
<td>bé</td>
<td>fon</td>
<td>o</td>
<td>phoneme</td>
</tr>
<tr>
<td>be</td>
<td>ɛ</td>
<td>be</td>
<td>font</td>
<td>ɔ</td>
<td>source</td>
</tr>
<tr>
<td>que</td>
<td>e</td>
<td>that</td>
<td>son</td>
<td>ɔ</td>
<td>sleep</td>
</tr>
<tr>
<td>quê</td>
<td>ɛ</td>
<td>what</td>
<td>són</td>
<td>o</td>
<td>be (3rd singular)</td>
</tr>
</tbody>
</table>