

ARTICLE



Visual reinforcement through digital zoom technology in FL pronunciation instruction

Siqi Wang, Shanghai University of Finance and Economics

Jian Li, Shanghai University of Finance and Economics

Qian Liang, Complutense University of Madrid

Abstract

Drawing on skill acquisition theory (DeKeyser, 2017) and the Information Feedforward and Feedback Loop model (de Bot, 1980), this study aimed to explore the effects of digital zoom technology as a visual reinforcement tool (VRT) in foreign language (FL) pronunciation instruction on learners' segmental production, and learners' attitudes toward and experience with it. The study was conducted during a two-week introductory FL Spanish course with a cross-over design. In the experimental class, the teacher used a tablet to provide magnified visual feedforward for articulatory gestures during explicit instruction of target consonants that were new to Chinese learners, and the students used smartphone apps with digital zoom for augmented visual self-feedback during their practice. In the control class, the teacher adopted a traditional analytic-linguistic approach for explicit instruction and an audio-only intuitive-imitative approach for students' practice. The results of the production test show that the experimental group performed significantly better in the production of the postvocalic /l/ and the dental fricatives /θ/ and /ð/, but not in the trill /r/. Interview data suggests that the VRT strategy was advantageous in directing the participants' attention to articulatory gestures, and that most students showed positive attitudes toward this new method.

Keywords: *Mobile-assisted Language Learning (MALL), Computer-assisted Pronunciation Training (CAPT), Pronunciation, Second Language Acquisition (SLA)*

Language(s) Learned in This Study: *Spanish*

APA Citation: Wang, S., Li, J., & Liang, Q. (2024). Visual reinforcement through digital zoom technology in FL pronunciation instruction. *Language Learning & Technology*, 28(1), 1–26.

<https://hdl.handle.net/10125/73558>

Introduction

Skill Acquisition Framework

Skill acquisition theory defines two types of knowledge, declarative knowledge and procedural knowledge, with the former being referred to as “knowledge that” and the latter “knowledge how” (DeKeyser, 2017, p. 16). Second language (L2) instruction mainly starts from the learning of declarative knowledge. Deliberate and repeated practice catalyzes the transition from “effortful use to more automatic use of the target language” (Lyster & Sato, 2013, p. 71). Three stages can be distinguished concerning the practice required for L2 learning: declarative knowledge, proceduralization and automatization (Pérez-Vidal, 2017).

Although “the large amount of practice necessary for the gradual reduction of reaction time, error rate, and minimal interference with other tasks that characterize the automatization process” (DeKeyser, 2007, p. 216) cannot occur in classroom instruction, in-class L2 teaching may offer repeated practice and feedback, which greatly contributes to proceduralization (Sato & Lyster, 2012). Moreover, declarative knowledge could not be easily proceduralized if processes like noticing are not promoted through

technical aids (Kennedy & Trofimovich, 2017). Through guided practice activities in tandem with corrective feedback, learners may achieve meaningful learning more rapidly and efficiently with a smaller error rate (Anderson & Schunn, 2000). This proves to be particularly true for the foreign language (FL) pronunciation learning process, whose objective is to achieve desirable performance in speech production with few opportunities to practice the target language outside the FL classroom (Dai & Wu, 2021; Saito & Lyster, 2012).

However, pronunciation has long been marginalized in FL instruction (Derwing & Munro, 2005), losing out to the more prominent vocabulary and grammar instruction in integrated FL courses. For example, Foote et al. (2016) reported that pronunciation accounted for only 10% of all the language-related episodes in an intensive English program in Quebec, Canada. To overcome the time limitation in teaching pronunciation, researchers and practitioners should focus on how to harness technological tools to optimize the proceduralization loop for the target sounds, involving production practices in conjunction with feedback (Lyster & Sato, 2013; Suzuki et al., 2019).

Information Feedforward and Feedback Loop in Teaching Pronunciation

As a complicated motor task, speech production requires co-work between feedforward and feedback subsystems (Guenther, Ghosh & Tourville, 2006). Different from general forms of feedforward and feedback, information feedforward (IFF) concerns the delivery of the objective set and the enhancement of the way in which that objective is attained, and information feedback (IFB) indicates whether the set objective has in fact been achieved, and why (or why not) (Bilodeau, 1966). Specifically, IFF and IFB often take the form of instruction, correction, and practice.

In response to the essential role played by explicit instruction and practice in skill acquisition, the current study proposes an IFF and IFB loop. According to Kröger et al. (2010), “phonetic treatment” in second language pronunciation learning asks for a great amount of “interaction between teacher and learner” (p. 338). First, the teacher must detect the students’ phonetic errors and make students aware of the problems; and second, the students must make timely corrections of the articulations. Feedforward for model imitation combined with feedback for error correction should be repeated “until these behaviors have become automatic and error-free” (VanPatten & Williams, 2015, p. 21). In other words, a number of such loops facilitate the unconscious automatization of declarative knowledge into the pronunciation skill required, because they can place both teachers and students in a real-time scenario and thus “close the gap between the current level of performance and the expected learning objective” (Koen et al., 2012, p. 240).

In pronunciation teaching, IFB is more often applied in a narrow sense, taking the form of extrinsic IFB, which is also referred to as “reinforcement” (de Bot, 1980, p. 36). De Bot (1980) further categorizes this IFB or reinforcement into auditory, visual, and tactile types, among which “auditory and visual IFB and IFF have been applied on a large scale” (p. 36) in the field of pronunciation teaching. However, research-oriented laboratory equipment only presents information that is “difficult for the layman to interpret” (p. 38), and in order to facilitate self-feedback and self-monitoring in educational settings, learner-friendly visual IFB and IFF has gained increasing attention. In practice, the IFF and IFB loop could be realized with the incorporation of visual reinforcement techniques into the pronunciation learning environment (Kröger et al., 2010).

Articulatory Visual Reinforcement

The notion of multisensory modes put forward by Celce-Murcia et al. (1996) highlighted the supplementation of the visual modality in FL pronunciation teaching. Traditional pronunciation instruction used to be supplemented by visual reinforcement mostly in the form of images and pictorial representations of phonemic symbols (e.g., vowel charts, phonemic charts, diagrams, etc.) (Offerman & Olson, 2016; Wrembel, 2007). Entering the new century, visual reinforcement techniques have gained considerable support with the application of modern technology. Inspired by the popularization of multimedia tools, some instructors of FL pronunciation have begun to employ multimedia learning aids, providing on-screen visual cues accompanying the audio soundtrack of a video (Bliss et al., 2018).

Generally, these technology-aided visual reinforcement techniques (VRT) fall into two paradigms (Kröger et al., 2010, pp. 339–340):

1. Acoustics-related aids: displaying auditory speech signals (e.g., oscillograms, spectrograms, contours).
2. Articulation-related aids: displaying vocal tract organs or articulators (e.g., lips, tongue, velum, larynx with glottis), showing the positioning and movement of articulators in sound.

Acoustics-related aids in computer-assisted pronunciation training (CAPT) have been used to help with learners' connected speech, for example, through providing visual representations of pitch/intonation contours (Chun et al., 2015; Levis & Pickering, 2004), waveforms, and spectrograms (Liu & Tseng, 2019; Patten & Edmonds, 2013). However, acoustic signals as such are difficult for students to interpret without expertise or guidance from the teacher (Rogerson-Revell, 2021). In principle, a visual display of speech sounds should “enable learners to process this visual information easily” and tend “not to discourage learners without any technical or scientific background” (Kröger et al., 2010, p. 339). Therefore, articulatory VRT that displays articulatory movements performed by either a model or the learner him/herself has begun to attract increasing interest in CAPT.

Both Motor Theory (Liberman & Whalen, 2000) and Direct-Realist Theory (Best & Tyler, 2007) assume that the representation of L2 speech information in the brain is based on relevant articulatory gestures (i.e., the way to use articulators, such as the tongue, lips, and jaw, to produce speech sounds) instead of acoustic signals. Visualizing articulatory gestures has been applied as an effective tool in speech production training, because “visual information would be used within the speech production network and supplying visual articulatory information to improve articulator placement and movement” (Haldin et al., 2018, p. 4). Massaro and Light (2003) employed a talking head (revealing the internal articulatory processes of oral cavity) as a tutor for Japanese learners' perception and production of L2 English /r/ and /l/ in a laboratory setting. They reported that no significant advantages were detected in inside articulation compared with mere frontal view. Then Li and Somlak (2017) applied pre-recorded videos of a tutor's visual articulatory gestures in teaching Chinese students' production of L2 English /θ/-/s/ and /ð/-/z/. And results of pre-, post- and delayed-tests showed significantly greater effectiveness of such audiovisual aids than audio-only instruction.

However, it is worth noting that these studies stress the articulatory gestures of the model only, which indicates that IFF regarding segmental production is offered separately during the pronunciation teaching process. According to the aforementioned importance of forming IFF and IFB loops, articulatory gestures related to the learners' own articulation in classroom-based FL pronunciation teaching also should not be neglected, as students need to be able to self-monitor and self-correct their speech sounds (Lipetz & Bernhardt, 2013) during their practice before receiving problem-targeted instruction from the teacher. Moreover, it has been suggested that in comparison with pre-recorded materials, real-time visual display boosts perception of auditory levels, leading to the acquisition of articulatory patterns that were not previously possible (Cleland et al., 2015).

Accordingly, there are two main concerns related to articulatory VRT, namely the degree of visibility achieved and the subject being visualized. First, in terms of visualizing real-person articulation, existing technology that could be used in pronunciation teaching mainly includes zoom technology and ultrasound imaging. Unlike ultrasonic tongue-imaging which only displays the tongue movements of the speaker, zoom technology provides a magnified frontal view of the vocal cavity through either optical zoom or digital zoom. Optical zoom is performed by changing between a set of zoom lenses with different focal lengths, and digital zoom by “selecting (cropping) a portion of the pixels in the camera lens and performing an interpolation to reach the desired image size” (Toutouchi & Izquierdo, 2018, p. 1). Obvious readability and hardware issues with both ultrasound imaging and optical zoom have made them unsuitable to meet the requirements for application in a classroom setting. Second, in the limited studies of articulatory VRT in an educational setting, no tools can satisfy the simultaneous need to display both the teacher's articulatory feedforward and the students' real-time self-feedback. Given the above issues,

there is a clear need for a convenient tool capable of supporting the IFF and IFB loops for visual articulatory gestures in L2 pronunciation teaching classroom.

The Study

In-app Digital Zoom Technology

This study adopts digital zoom technology, the method of choice in the cameras in modern smartphones and tablets. As described above, digital zoom yields an overall augmented view of the subject not by moving lenses but by processing the image captured, which is realized through software and algorithms built into photography applications on mobile devices. In fact, Glaser et al. (2017) have already demonstrated that digital zoom can be used to direct learners' attention to important elements in visual learning material. Thus, the present study attempts to apply this technology to FL pronunciation teaching. Modern mobile devices are equipped with two sets of cameras. The rear-facing camera can serve the function of augmenting the teacher's articulatory IFF and the front one the students' real-time information self-feedback. For the sake of visibility among the whole class, we suggested that a tablet with a larger screen than a normal smart phone should be assigned to the teacher to provide visual feedforward. And to zoom in to the most degree while maintaining a complete mouth view, the rear lens of the iPad mini used in this study was positioned about 28–31cm away from the teacher's mouth (see [Figure 1](#)). Mobile phones were used by the students for real-time self-feedback.

Figure 1

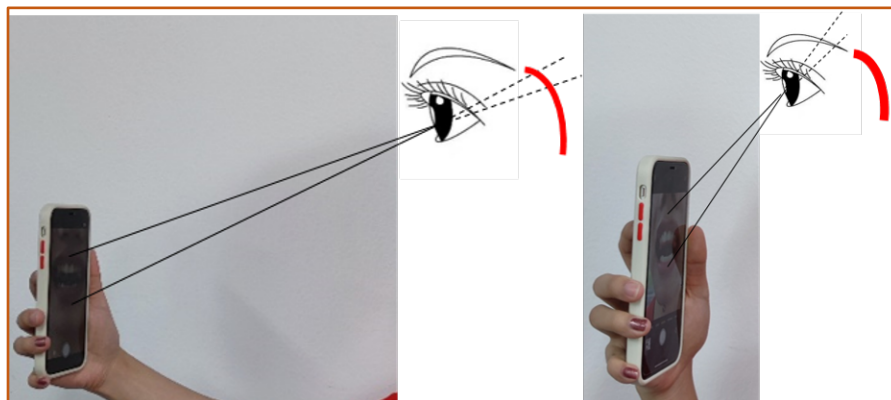
Articulatory Feedforward Given by a Teacher Using the Rear-facing Camera of a Tablet



In terms of providing real-time self-feedback, mirrors have been widely used as a useful visual aid to improve the speech of both the hard of hearing and L2 learners with normal hearing (Celce-Murcia et al., 1996; de Bot, 1980). However, the use of mirrors has a fatal flaw: the reflection in the mirror is not magnified, and as a result, the visual information related to certain articulatory gestures cannot be clearly captured. This is no different from a front-facing camera without zoom technology. When monitoring articulatory gestures with a mirror or a front camera, the user has to move the device extremely close to their face so that the image becomes inaccessible to their eyes (see [Figure 2](#) right). However, with the digital zoom built into the in-app camera we can obtain a magnified close-up view of the same magnification level from a more comfortable distance to suit our eyesight (see [Figure 2](#) left).

Figure 2

Comparison of Eyesight for the Same Close-up Views Using the in-app WeChat Camera with Digital Zoom (left) and the No-digital-zoom System Camera (right)



Note. The red curve represents the retina.

Importantly, there is more to the use of the front camera than meets the eye. When using the rear camera as described above, we simply open the system camera application on the device and zoom in on the image to obtain the ideal effect because there is already a digital zoom algorithm built into the camera. However, this is not the case for the front camera. Many people may not notice that the selfie mode in the system camera application on any smart phones or tablets does not allow zooming, but third-party selfie applications and many social media applications with access to the system camera do. Their obvious difference is illustrated below in Figure 3. Based on this technological premise, this study employed the in-app camera function built into WeChat as the experimental instruction condition for students' practice.

Figure 3

Comparison Between Frontal Views of Mouth by System Camera with no Digital Zoom (left) and by In-app WeChat Camera with Digital Zoom (right)



Target Segments

Previous studies have shown that Chinese learners of Spanish have more problems with Spanish consonants than vowels (Chen, 2017; Lu, 1991). After analyzing the speech production of 51 Chinese college students in a Spanish beginner course, Chen (2011) found that the multiple alveolar trill /r/ created the biggest barrier for Chinese novice learners, and half of the participants mispronounced the postvocalic

/l/ as the retroflex lateral /ɭ/. An empirical study by Fernández (2015) detected the following consonant problems encountered by Chinese learners of Spanish: voiced /b/, /d/ and /g/; alveolar lateral /l/ especially in the postvocalic position; alveolar multiple trill /r/; and dental fricatives /θ/ and /ð/.

Second Language Acquisition (SLA) research (e.g., Flege et al., 2021) has investigated how cross-language mapping patterns may interfere with the acquisition of segments. The Speech Learning Model (SLM) developed by Flege (1995) and the Perception Assimilation Model (PAM) described by Best (1995) both claim that the most difficult sounds are those that are the most similar (but not identical) to the L1 sounds. With a focus on the construction of a new phonetic category, SLM and the revised version (SLM-r) developed by Flege et al. (2021) point out that for those sounds that are nonexistent in L1, learners can simply construct a new phonetic category. However, L2 learners cannot do this for L2 sounds that are similar to L1 counterparts without proper training, since they often have difficulty in perceiving their fine-grained phonetic differences. The PAM (Best, 1995) and its L2 extension (Best & Tyler, 2007) adopt the lens of articulatory gestures and predict that L2 learners cannot perceive the distinction between two L2 sounds if they are both assimilated to a single L1 category, while L2 sounds that are articulatorily distinct from L1 sounds are ultimately easier to learn.

Based on the connection between articulatory similarity and difficulty for L2 learners predicted by the major L2 speech theories, we conducted a detailed cross-linguistic articulatory analysis of the target segments.

Table 1

Cross-linguistic Analysis of the Target Spanish Consonants in terms of Articulatory Classification

Segment	Manner of articulation	Place of articulation	Phonotactic constraints
/θ/ - /ð/	+	–	N/A
/l/	+	+	–
/r/	–	+	N/A

Note. “+”: in Chinese Mandarin; “–”: not in Chinese Mandarin.

The Spanish dental fricatives /θ/ and /ð/ are incorrectly replaced by the articulatorily-similar alveolar /s/ and /d/ in Chinese students’ production (Fernández, 2015), because, as shown in Table 1, there are no dental consonants in the Chinese phonetic inventory with respect to the place of articulation. As a result, the dental place of articulation creates the main obstacle for Chinese learners. The case of /r/ is just the opposite: in the Chinese Mandarin consonant inventory there are many phonemes with the same alveolar place of articulation as /r/, but the trill manner of articulation is absent in most Chinese language families.¹ Therefore, Chinese learners tend to replace /r/ with /l/, which is perceived as being articulatorily similar and whose place of articulation is identical to /r/.

The reason why the post-vocalic /l/ stands in the way of Chinese L2 learners of Spanish in terms of their successful pronunciation performance is that it appears in a different phonetic environment from their L1. In Spanish the segment can appear in both postvocalic and prevocalic positions, while in Chinese it only appears before a vowel. Chen (2011) found that Chinese learners of Spanish often replaced the postvocalic /l/ with the retroflex lateral /ɭ/ or /ʎ/, whose place of articulation is more familiar. Deterding (2007) concluded that Chinese students of English tend to vocalize the postvocalic /l/ into a back vowel /ʊ/, or may even delete it when it appears after a back vowel /ʊ/. Zhu et al. (2022) reported that for English learners of intermediate and advanced levels in the Chinese university without specific phonetic training, the accuracy rate of producing postvocalic /l/ was almost zero. This shows that although the postvocalic /l/ is also used in English, the prior knowledge of English does not contribute much to its acquisition in Spanish if the learner has not taken any phonetic courses before.

To test the potential of digital zoom technology in facilitating Spanish segmental production, the study focused on the trill /r/, the postvocalic /l/, and the dental fricatives /θ/ and /ð/ as target segments, because they involve more observable articulatory gestures than other problematic sounds such as /b/, /d/ and /g/, whose voicing features also cause great pronunciation hardship for Chinese learners but which involves less visible articulators (e.g., vocal folds).

Purpose and Research Questions

This study sets out to explore the feasibility and effects of digital-zoom-aided VRT in a FL Spanish pronunciation class in terms of targeted consonant production achievement. Furthermore, since previous studies have revealed that learners hold different attitudes toward mobile-assisted language learning strategies (Seibert Hanson & Brown, 2019; Wu, 2021), it is necessary for us to consider the participants' perceptions of the VRT strategy used in this study. Therefore, the present study aimed to answer the following two research questions:

RQ1: To what extent, if any, does VRT aided by digital zoom technology facilitate the participants' targeted segmental production?

RQ 2: What is the participants' attitude toward the new VRT strategy?

Methods

To address the above two research questions, this study adopted a mixed-methods design, including the following quantitative and qualitative methods: 1) cross-over experiment; and 2) semi-structured interview.

Cross-over Quasi-experiment

Participants

The participants were 30 students who were novice learners of Spanish, recruited regardless of their majors from a university in Shanghai, China (24 females and 6 males). They were randomly divided into two groups, namely Group 1 ($N_1 = 15$, mean age = 22.2) and Group 2 ($N_2 = 15$, mean age = 21.75). The two groups participated in the study as two classes in different time slots. Every participant received 50 RMB as compensation for taking part, and gave his or her written informed consent for participation and the use of data for research purposes.

Prior to the study, the students were asked to complete a background survey. The survey showed that none of them had any previous experience of Spanish learning, and none had ever taken any phonetic courses. The 30 participants were from 20 provinces across the country.

Target Sounds

The target items were the Spanish trill /r/ in the absolute initial position, intervocalically, and in the post-consonant position, the postvocalic /l/, and the dental fricatives /θ/ and /ð/ in the syllable initial and coda positions.

Course Context

Each group received instruction once a week for a total of 2 sessions, with each session lasting 45 minutes. The sessions simulated lessons in the optional and introductory Spanish course that was offered to college students taking non-Spanish majors, aiming to help them acquire basic Spanish competence in speaking, listening, reading, and writing. The course context has been widely applied to introductory teaching of FL other than English in China's higher education curriculum (Chen et al., 2021). To be specific, the course content is composed of: a) a brief introduction to Spanish and the alphabet; b) teaching pronunciation of all the five vowels; c) teaching pronunciation of 13 consonants (target segments included); d) explaining the rules of syllable segmenting and stress; and e) basic conversational expressions. The four target sounds were divided into two groups (/r/-/l/, /θ/- /ð/), for the sake of the cross-over design and teaching

convenience. Each group of sounds was instructed over one session. The teacher had been teaching L2 Spanish for seven years with a Master's degree in linguistics, but had not researched phonology. This means that the course was not weighted toward pronunciation input; this is important because the instructor's research focus has been found to influence the input that they provide during class (Long, 2017).

Testing Materials

The production tests employed read-aloud tasks carried out on desktop computers, in which 15 words containing the target sounds (/r/-/l/ in production test 1, /θ/- /ð/ in test 2) and five fillers were shown on the screen for students to read. Since the study is quasi-experimental and targeted at novice learners, it was unfeasible to employ "controlled and spontaneous speech elicitation tasks" which would seem to be more suitable to measure the extent of unconscious proceduralization (Saito & Plonsky, 2019, p. 665). However, we tried to control the planning time during the test as each word was only displayed for five seconds before the screen automatically transitioned to the next one.

The participants were told to read the word once only the moment it appeared; repetition and correction were not allowed. All the possible phonetic environments for each consonant taught during the sessions were covered. The testing stimuli were different from those used in the teaching sessions. Stimuli used in teaching and testing can be found in the [Appendices A and B](#).

Apparatus

The study was conducted in an audio-visual room used for listening and speaking classes, where every student has a desktop computer equipped with a headset and microphone. During the experimental sessions the teacher used an iPad mini positioned 28–31cm in front of her mouth to magnify the articulatory gestures. Then in the practice part the students were asked to practice their articulatory gestures using their smartphones.

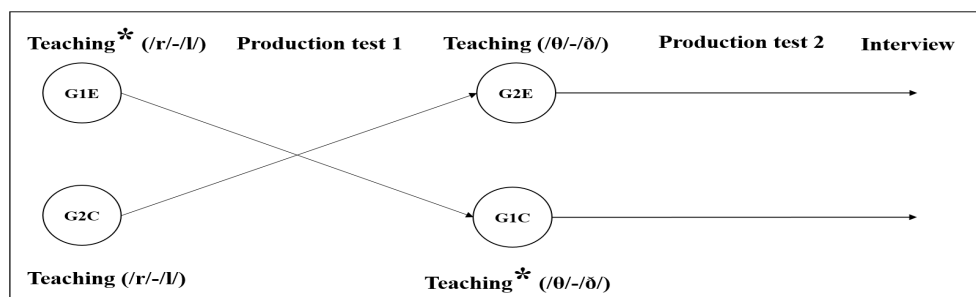
Study Design and Procedure

Cross-over Design

We chose a cross-over design because of several advantages offered by this type of study: first, both groups gain access to the training; second, the intervention could be tested with a relatively larger sample; and third, participants in both groups are exposed to similar amounts of extra training (Fouz-González, 2020). Participants were randomly divided into two groups (G1 = 15, G2 = 15). First those in G1 acted as the experimental group (G1E), with G2 as the control group (G2C). After production test 1, G2 were taught with the technology-aided strategy (the experimental condition) and thus transferred to the experimental group (G2E), while G1 was put under the control condition (G1C). After the second session both G2E and G1C took production test 2 ([Figure 4](#)).

Figure 4

Cross-over Design of the Study



Note. *with digital zoom intervention.

Teaching Procedure

Each teaching session lasted 45 minutes and consisted of an instruction part and a practice part. In the teaching session in the first week, after a brief overview of Spanish, the teacher taught the alphabet, the five vowels, the consonants /m/, /p/, /b/, /f/, /r/, and /l/, as well as the corresponding spelling rules. In the second session, /n/, /k/, /g/, /t/, /d/, /s/, /θ/, and /ð/ were taught and practiced for the first 35 minutes. The last 10 minutes of both sessions were for simple conversational expressions in Spanish, such as basic greetings and introductions (Figure 5). Most of the time for pronunciation training was spent on the target segments which presented the biggest challenges to the students. Specifically, during each experimental instruction, students observed the teacher's demonstration with the iPad mini for about 10 minutes and practiced with the in-app digital zoom on their smart phones for 15 minutes.

Figure 5

Screenshot of a Teaching Slide Used for Basic Spanish Conversation

¡Hola!	你好!	English translation: Hello!
¡Buenos días!	早上好!	Good morning!
¡Chao!	再见!	Bye!
¿Cómo te llamas?	你叫什么名字?	What's your name?
Soy...	我叫.....	I am...
¡Mucho gusto!	很高兴认识你(您)!	Nice to meet you!
--¡Hola! ¡Buenos días! ¿Cómo te llamas?		-Hello! Good morning! What is your name?
--¡Buenos días! Soy Ana. ¿y tú?		-Good morning! I am Ana, and you?
--Soy Mario. ¡Mucho gusto!		-I am Mario. Nice to meet you!
--¡Mucho gusto!		-Nice to meet you!

In the instruction related to the target segments, the teacher first taught using the analytic-linguistic approach for both groups, in which she explained explicit articulatory knowledge about the target sounds with detailed articulatory descriptions shown on the slides (Figure 6). Then, she demonstrated with her mouth to the class without any support from tools or technology for the control group, while for the experimental group she demonstrated the articulation using the rear-facing camera of an iPad.

Figure 6

Screenshot of a PPT Slide Showing Articulatory Knowledge of the Target Trill Sound

Translation : Pronunciation of the consonant r. Multiple trill: when the letter r is the onset or appears after 'n, l, s' within a word, or written as 'rr', it is a multiple alveolar trill. When you pronounce it, lift up the tip of your tongue and try to relax the blade to let the airstream through, and then vibrate the vocal cords to make the tongue tip trill multiple times.

*辅音字母 r 的发音

(1) 多颤: 当字母r在词首或词内的n、l、s后面, 以及在词中拼写成rr时, 为舌尖齿龈多击颤辅音。发音时, 舌尖抬起, 舌部尽量放松, 然后让气流通过, 在声带振动的同时, 舌尖多次颤动。

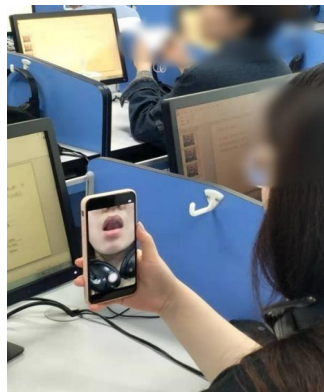
ra	re	ri	ro	ru
arra	erre	irri	orro	urru

rata	reno	rito	rosa	rutina
enrase	Enrique	sonrisa	alrededor	las risas
torre	perro	narrativa	carreta	farruco

During the practice part, the students in the control group were told to read the words from the textbook after the teacher. They imitated the sounds merely using their auditory competence. Students in the experimental group were asked to practice using the “camera” function in the chat-box of the WeChat supported by digital zoom, while watching the teacher’s visual IFF (see Figure 7). After each teaching session a production test was administered. It must be pointed out that the current study did not arrange pre-tests, considering that it would be difficult for absolute beginner participants to fulfill the read-aloud task without any knowledge of Spanish sound-spelling relations.

Figure 7

Students Practicing Using Digital Zoom



Evaluation

Production Test

The participants’ productions were evaluated by three native Spanish experts in terms of accuracy. A fourth judge was asked to disambiguate in cases of disagreement. The experts were all experienced Spanish teachers who had taught Spanish for over 8 years. Two of them held PhD degrees in applied linguistics, and the other two had majored in FL education.

The ratings were dichotomous (1 if the target sound was pronounced adequately, 0 if it was mispronounced). The raters could play each stimulus as many times as they needed. Interrater reliability was measured using Fleiss Kappa, with the test yielding a Kappa value of 0.781, which can be interpreted as substantial agreement (within the range 0.61–0.8). Intra-rater reliability was assessed by analyzing the raters' consistency in rating 12 extra items that had already been rated (three per target sound). No instances were found in which raters assigned a different score to an item that had already been evaluated, so no extra tests were conducted.

Interview

After the experiment, a brief sampling interview was administered to investigate the participants' attitudes toward the technology-aided teaching approach. Six participants from both groups were chosen according to their test scores when they acted as the experimental, among which there were five scoring higher than 0.8, two average (equal or less than 0.5). One-on-one semi-structured interviews were carried out in Mandarin Chinese, with each interview lasting 20 minutes (see [Appendix C](#) for interview questions). The interviews were recorded with the permission of the participants and were later transcribed for qualitative analysis using a thematic approach.

Results

Production Test

This test was administered to investigate whether VRT aided by digital zoom technology could facilitate the participants' targeted segmental production. The accuracy percentage (i.e., number of "1" scores / total amount of words) for each participant was calculated based on ratings. The scores ranged from 0 to 1.

An independent t-test (see [Table 2](#)) showed a significant difference between the overall scores achieved by G1E and G2C ($t = 3.344$, $p < .05$, effect size of Cohen's $d = 1.22$) in the teaching of the first group of sounds, which meant that the overall test scores for G1E's production were significantly higher than those of G2C ($MD = 0.27$).

Table 2

Test Results for Overall Accuracy Percentage of /r/ and /l/

Group	<i>N</i>	<i>Mean</i>	<i>SD</i>	<i>t</i>	<i>p</i>	<i>MD</i>	Effect size
G1E	15	.593	.231				
G2C	15	.323	.211	3.344	.002*	.27	1.22

Note. * $p < 0.05$

When we further examined the accuracy percentages for /r/ and /l/ separately, as shown in [Table 3](#), no statistically significant effect was found for the test scores of the trill /r/ ($t = 1.934$, $p = 0.063$), but G1E's test scores for the production of postvocalic /l/ (/l/: $t = 2.394$, $p < .05$, effect size = 0.87) were significantly better than those of G2C ($MD = 0.293$).

Table 3

Separate Test Results for Accuracy Percentage for /r/ and /l/

Target	Group	<i>N</i>	<i>Mean</i>	<i>SD</i>	<i>t</i>	<i>p</i>	<i>MD</i>	Effect size
/r/	G1E	15	.447	.374	1.934	.063	.247	0.71
	G2C	15	.200	.322				
/l/	G1E	15	.740	.261	2.394	.025*	.2933	0.87
	G2C	15	.447	.396				

Note. * $p < 0.05$

Therefore, the VRT aided by digital zoom did facilitate the experimental group in their production of /l/, with a significant increase in their accuracy percentage compared with that of the control group, but this was not found for the trill /r/.

The results for the second group of sounds (/θ/ and /ð/) are given in Table 4, showing that G2E performed significantly better than G1C in the production test ($t = 2.0857$, $p < 0.05$, $MD = 0.14$, effect size = 0.76), in terms of overall scores.

Table 4

Test Results of Accuracy Percentage for /θ/ and /ð/

Group	<i>N</i>	<i>Mean</i>	<i>SD</i>	<i>t</i>	<i>p</i>	<i>MD</i>	Effect size
G2E	15	.713	.192	2.0857	0.046*	0.14	0.76
G1C	15	.577	.166				

Note. * $p < 0.05$

No separate inferential analysis was conducted concerning the fricatives because the two sounds only differed in voicing, which is not salient for visualization and does not fall into our study focus.

Interview

The results of the qualitative analyses, addressing the second research question, yielded three major themes: technical, social/psychological, and academic (Dai & Wu, 2021). In the following paragraphs extracts are used to illustrate the students' attitudes toward the new VRT strategy. The six interviewees (two males and four females) are identified by pseudonyms as M1, M2, F1, F2, F3, and F4.

Technical Affordances

All the interviewees appreciated the technical affordances of the digital zoom camera used in the teacher's IFF and in the IFB during their practice. The top feature that all of them mentioned was the greater visual clarity that the technology provides: "*The magnified function of the digital camera has made the pronunciation more graphic to us*" (F1); "*The clear and intuitive way of displaying the articulators is much better than listen-and-repeat teaching method*" (F2). The second attractive feature of the technical affordances was convenience. All the interviewees reported that it was convenient to use the digital camera built into social or selfie apps in their class practice, since smartphones have become an integral part of modern life. The two male interviewees said they would have refused to carry mirrors with them even if they were required to use them in pronunciation classrooms, although the female interviewees found this acceptable.

Two of the interviewees also mentioned that the use of a tablet facilitated the real model presentation by

the teacher, which they preferred to pre-recorded videos:

“She used an iPad to magnify her mouth movements. This is more effective than videos and animations because it is real person modeling, and we have got a clear view of how the sound is produced” (F2); “I like the real-person demonstration ... because the video pre-recorded by native Spanish speakers lacks 3-D feeling.” (F3)

Academic Gains in Production of the Target Segments and Articulatory Attention

No matter how they performed in the production tests, all the participants reported that the VRT strategy benefited their production of the target Spanish sounds in general except for the trill /r/ which is particularly difficult for Chinese students, as some of them complained. Only one participant mentioned that the VRT strategy helped her master the trill sound: *“The biggest gain for me is that I have learned to produce the /r/ sound, which seemed difficult for most people” (F3)*. Furthermore, the participants provided their own explanations for how it benefited their pronunciation learning: *“I learned more about phonetics in this way which added to my understanding of phonetic knowledge” (M1)*. Obviously, practice aided by VRT not only served to proceduralize the participants’ declarative phonetic knowledge, but also helped them reassess and restructure inaccurate prior knowledge (Lyster & Sato, 2013).

Another academic advantage that most interviewees reported was the enhancement of their articulatory awareness. The articulatory-based paradigm aided by VRT contributed to the participants’ attention to the classification of speech sounds and their articulatory gestures while producing the target sounds, helping them to understand the rules regarding articulatory phonetics. The interviews with F2 and M1 indicated that they found that the traditional “listen and repeat” mode was insufficient for the teaching of all the phonemes, and supplementation of the visual articulation process was essential:

“I remember vowels are easier to master, but certain consonants could be very tricky because the position where you put your tongue and the extent to which you open your mouth have much to do with the sound you produce. Therefore, it is relatively hard to become aware of this if you simply listen to your own pronunciation” (F2);

“The advantage of viewing the teacher’s and our mouth movements is that it can make the students realize that the sound could not be pronounced merely with audio input, but also with correct movements of lips and tongues. Thus, demonstration of these movements becomes extremely important in pronunciation” (M1);

“(After using the digital camera), I became more aware of the articulatory gestures, such as the place of tongue and the shape of lips... For example, I’d never paid attention to the place of my tongue tip before when producing the postvocalic /l/ sound” (F4).

Through watching the augmented articulatory gestures made by the teacher and themselves, they came to recognize that speech production was not only about the sound; mouth movements are equally important. As well as audio input, visual reinforcement seems to play an indispensable role in pronunciation training. By using the VRT strategy the participants recognized that they had begun to take notice of the importance of articulatory gestures in segmental production:

“For languages whose sound systems are very different from that of our native language, like Spanish, this kind of demonstration and self-monitoring, including the mouth shape and tongue position, seem very crucial in producing these sounds and it is difficult for us to get the knack merely by listening and repeating on our own” (F3).

According to the self-report data, the VRT strategy aided by digital zoom was effective in directing the learners’ attention to articulatory gestures for those sounds that went against their old articulatory habits during their practice: *“I used not to notice if my tongue tip stuck out of my teeth while producing the sounds /θ/ and /ð/... I will pay more attention to this gesture” (M1).*

Social/Psychological Affordances

In terms of psychological affordances, two interviewees reported that the use of tablets and mobile phones brought extra enjoyment to their otherwise boring practice: *“I think this new method made the pronunciation teaching very interesting”* (F2). This participant was candid about the potential psychological pressure that practice with visual self-feedback might bring, which could be compensated by the psychological benefit of the IFF and IFB loop aided by VRT:

“If students were only asked to practice and monitor our own performance with the magnified camera built into the phone, I would feel under stress and maybe even embarrassed... However, with the teacher using an iPad to show her enlarged articulators to us, I feel quite relieved” (F2).

Furthermore, from the participants’ self-report interviews, we can see that the VRT tools were also conducive to ensuring high motivation during the learners’ classroom pronunciation practice activities and after-class practice routines: *“I am more willing to get involved in the practice activities, endeavoring to find out my pronunciation problems... I will adopt this type of practice after class”* (F1); *“I am going to stick to this practice with the digital camera as much as possible”* (F4).

With respect to social affordances, four of the interviewees reported that the IFF and IFB loop with VRT support provided them with more teacher-student interaction and connection, enhancing their class engagement: *“I feel more involved (in the class practice) ...”* (F1); *“This kind of practice made me feel more connected with the teacher... then I can find my pronunciation problems more quickly by comparing my articulatory gestures with the teacher’s”* (M1).

Concerns and Possible Solutions

Almost no interviewees reported any technical problems with the VRT use in class. One participant found it slightly distracting to look at her own mouth while producing target phonemes: *“It (looking at the phone) is useful to distinguish between similar phonemes, but it distracted me a little bit and I then lost attention to the sound I produced”* (F1).

F1’s remarks suggest one issue that learners do need to be aware of when using a VRT strategy, namely the coordination of multisensory modalities. A plausible reason for this difficulty is that learners are used to the traditional single-modality mode in pronunciation training, and practice is needed in order to be able to engage with more than one sensory modality: *“Two sessions were not actually enough for us to learn all the Spanish sounds. I wish we could take more practice in this new way with the model display (augmented feedforward) from the teacher”* (F3); *“The time allocated to watch and practice is limited in the class, and we need more time to get used to this and master the correct articulation”* (F4).

The practice time limitation reflected in the interview data offers a possible explanation for why the production test result for the trill /r/ was not significant for G1E: *“I can’t master the trill manner, even with the teacher’s magnified model when producing the Spanish /r/”* (F4); *“I probably need more time to practice the difficult /r/”* (F1). It has been reported elsewhere that the trill can cause great difficulty for Chinese learners (Chen, 2011), and producing the sound requires plenty of time spent in teaching and practice. Thus, it is hard to achieve successful production in only two weeks.

With respect to social concerns, one of the interviewees suggested that the IFB and IFF loop with VRT could also be introduced in peer feedback activities, with the aim of increasing peer interaction as well as engagement.

Interestingly, some of the participants contributed their ideas for how to further develop the VRT: *“I wish our practice process could be supported by more advanced technology like real-time visual feedback with automatic error annotations, or real-time teacher visual feedforward with the analysis of sagittal section ... Ha ha! maybe too imaginative...”* (F3).

Discussion and Implications

The current study has addressed two research questions derived from several gaps in the existing literature, specifically related to how digital zoom as a VRT used in an IFF and IFB loop can facilitate the proceduralization of articulatory knowledge during the acquisition of segmental production skills in FL classrooms. The characteristics of our research context need to be further clarified before we offer a more focused analysis of our quantitative and qualitative results.

First, as mentioned in the [Introduction](#), pronunciation instruction is usually compressed into a very short period of a FL teaching course (Foote et al., 2016), especially for elective courses in higher education contexts. Second, the L2 Chinese students of Spanish in this study were all adult starters; compared with child starters, they are less likely to perceive the phonetic and articulatory distinctions between a L1 sound and a similar but not identical L2 sound (Hazan et al., 2005; Walley et al., 1993). Third, our FL participants were able to rely only on classroom instruction and practice with conscious use of rules, having few opportunities to practice in natural environments compared with L2 students who are able to study abroad (Pérez-Vidal, 2017). Last but not least, adult FL learners usually experience language anxiety in pronunciation classes related to receiving “theoretical knowledge, straightforward assessment and performing in front of others” (Khoroshilova, 2016, p. 541), which may prevent them from making noticeable progress in FL pronunciation learning (Baran-Lucarz, 2013). This is especially true of Chinese adult students (Liu, 2006). These deficiencies make it more difficult for FL teachers to help the students proceduralize declarative knowledge during limited in-class practice sessions.

Our results show that the use of VRT can enhance the efficiency of pronunciation practice in adult FL classrooms from the following three aspects: segmental production gains, articulatory awareness, and social/psychological benefits. Although the results of the production tests showed a significantly facilitative effect of the VRT strategy on the participants’ overall production accuracy for the two groups of target consonants, their performance in the production accuracy of the four targeted phonemes varied. The experimental group outperformed the control group in the production tests for the interdental fricatives /θ/ and /ð/ ($p = 0.046$) and for postvocalic /l/ ($p = 0.025$), but no statistically significant effect was found for the test scores of the trill /r/ ($p = 0.063$). Obviously, the new VRT strategy worked better in helping the students master the production of postvocalic /l/ and the dental fricatives /θ/ and /ð/, while improvements in the production accuracy of the trill /r/ were less satisfactory.

Combining the cross-linguistic articulatory analysis of the target segments (shown in [Table 1](#)) and the production test results for postvocalic /l/, /θ/, and /ð/ (shown in [Tables 3](#) and [4](#)), we see that the VRT strategy has more effect on the establishment of declarative knowledge about the place of articulation than about the manner of articulation. Place of articulation consists of knowledge about which articulators are involved and where the tongue should be put, while manner of articulation means how the sound is made through tongue movement, which is obviously more complex and is difficult to observe without an intrusive camera. Postvocalic /l/, /θ/, and /ð/ in this case just requires the visual concentration of articulatory information of the tongue apex, but the trill /r/ involves the complex movement of tongue posterior sections (del Puy Ciriza & Rivera-Campos, 2020). It is evident that the VRT in this study was much less successful in helping them with the difficult trill production which is associated with motoric learning in the short term. The interviews confirmed that the participants felt they needed longer practice with the VRT aids in order to master the difficult trill /r/. These results also partially support Li and Somlak’s (2017) and Hazan et al. (2005)’s findings that visual-audio training is only an advantage in training the production of sounds with sufficiently noticeable articulatory gestures, and will not work for all L2 sounds.

The results of the semi-structured interviews lend support to the Noticing Hypothesis, which claims that conscious attention to form in the input or in the feedback that learners receive is necessary for learning to take place (Schmidt, 1993). The participants reported that their attention was successfully guided to the problematic L2 articulatory features with the help of the digital zoom, and that they began to reflect on

the explicit knowledge related to articulatory gestures that the teacher delivered. This means that students engaged in some degree of metalinguistic reflection, which might have helped them to become more aware of explicit or declarative knowledge. As DeKeyser (1997) argues, explicit declarative knowledge lays the foundation for repeated practice, which facilitates the first stages of proceduralization—that is, practice with conscious use of articulatory knowledge. This finding is also supported by the cueing principle from multimedia learning theory (Mayer 2009, 2014), which suggests that multimedia learning material with more cues added is more effective, because the cues help learners to select, organize, and integrate information (van Gog, 2014) by guiding their attention to specific learning elements. With production tests outcomes and FL in-class instruction, our findings strengthen and extend Glaser et al. (2017)'s study, which provided eye-tracking evidence that digital zoom may function as a procedural cue for increased visual attention, but which failed to examine learning outcomes.

Our VRT used in the IFF and IFB loop helped to cue the students' attention not only to the input from the instructor's model delivery, but also to their own output, which, as Swain (1995) points out, has a facilitating role in SLA, since "in producing the target language, ... learners may notice a gap between what they want to say and what they can say" (pp. 125–126) —in this case, the gap between the teacher's IFF and their own performances—promoting more vigilant monitoring of their output. Also supported by de Bot's (1980) IFB and IFF loop model which functions in the form of instruction and correction, the teacher's explicit instruction and the students' practice during the pronunciation teaching in this study achieved a high degree of unity and integration by promoting the noticing process. Students were able to grasp and close gaps between their own performance and the desirable one more quickly and easily. Hence, the two parts are no longer dichotomized, and the effects of Mobile-assisted Language Learning (MALL) on L2 skill acquisition can be maximized.

This synergy between the IFF and IFB loop and VRT aids is not confined to their overall role in directing the participants' attention to articulatory knowledge and leading them to change their habits during in- and after-class pronunciation practice. As revealed in the interview, the simultaneous use of VRT in both IFF and IFB also yielded social and psychological benefits for the learners. It increased teacher-student interaction and reduced anxiety and embarrassment related to error correction. The VRT used in our IFF and IFB loops encouraged synchronous interaction and communication, which is more favored by FL learners than asynchronous video input by model speakers. Moreover, the use of electronic tools in class instilled an element of fun into the otherwise dull and repetitive practice routine. This also helped to reduce the anxiety related to starting to learn a new FL: *"I know that learning a new FL requires a lot of time to practice. This new learning method (with the aid of VRT) helps to keep my motivation"* (F2).

In light of previous findings, our study suggests three pedagogical implications for mobile-assisted FL pronunciation instruction within the articulation-based paradigm. First, when adopting a VRT strategy for pronunciation teaching and considering the allocation of practice time for each target segment, instructors should carry out a cross-linguistic analysis of their place and manner of articulation, because the non-invasive VRT has a little effect in the short term on the motoric learning of the segments involving the movement of tongue posterior. Second, the audio-visual mode of practice supported by VRT, compared with the traditional listen-and-repeat mode, may not only contribute to the enhancement of FL learners' attention to articulatory knowledge, but also foster their awareness of monitoring the gaps between their own output and the teacher's model input. Third, when a new MALL tool is adopted for pronunciation FL instruction, teacher-student interaction and language anxiety should be attended to. Otherwise, the learning achievement in a MALL environment may be greatly affected, since higher levels of acceptance are closely linked with better learning achievement (Cheng & Chen, 2022; Sung et al., 2017).

Conclusion

This study has compared FL segmental pronunciation instruction supported by a VRT strategy via digital zoom technology with traditional instruction stressing the listen-and-repeat mode of practice without any visual reinforcement. The findings indicate that the experimental groups (G1E and G2E) showed better

performance in the production of postvocalic /l/ and the interdental /θ/ and /ð/, but not of the trill /r/. The interview data revealed that the VRT strategy was advantageous in cueing the students' attention to the articulatory gestures in the instructor's model delivery and their own output in the practice in tandem with self CF. The study also showed that the simultaneous use of digital zoom technology in the IFF and IFB loop was welcomed by the students, providing them with social and psychological benefits, such as a sense of dyadic interaction, belonging, and relaxation.

Based on these findings, we propose that articulation-based VRT should be adopted in FL segmental pronunciation instruction to enhance the efficacy of explicit articulatory knowledge delivery and deliberate practice in class. The VRT condition may reinforce the learning environment by providing plentiful, clear, and immediate model input, and students can better self-monitor their articulatory gestures and receive consistent feedback, instilling a greater sense of self-monitoring and autonomy in FL learners (Heift & Chapelle, 2012). It is hoped that this strategy will be widely integrated with more advanced mobile-based augmented reality (Zhu et al., 2022) or virtual reality technology in the future, as a good remedy to the abovementioned deficiencies inherent in adult beginner-level FL courses.

It should be noted that this study has at least two limitations. The first is the small sample size, which was due to the limited availability of participants. The other lies in the qualitative method used to measure the attention aspect of articulatory awareness. If we had used a questionnaire based on well-designed rubrics or think-aloud protocols as in Wrembel (2013), the quantitative data thus elicited might have been more significant in assessing this VRT's function of promoting noticing.

Acknowledgements

We would like to show our gratitude to the participants and judges. Also, we'd like to thank the anonymous reviewers for their valuable suggestions that have made the paper stronger. This work was supported by the National Social Science Fund of China [Grant No. 23BYY163].

Notes

1. Although the trill /r/ does exist in in some of the minority languages in China, such as Uyghur and Mongolian (Li, 2008; Chimitdorzhieva, 2018), not many native speakers of these languages choose to learn Spanish, since Spanish would be their fourth language behind their native language, standard Mandarin, and English. It is known that the trill /r/ exists in a few dialects in the central and northern regions of Hubei and southern Henan, but Li (1984) claims that the trill /r/ is tending to disappear, being affected by the spread of Putonghua (standard Mandarin). This is especially true of young people.

References

- Anderson, J. R., & Schunn, C. D. (2000). Implications of the ACT-R learning theory: No magic bullets. In R. Glaser (Ed.), *Advances in instructional psychology* (pp. 1–33). Lawrence Erlbaum Associates Publishers.
- Baran-Lucarz, M. (2013). Phonetics learning anxiety – results of a preliminary study. *Research in Language*, 11(1), 57–79. <https://doi.org/10.2478/v10015-012-0005-9>
- Best, C. T. (1995). A direct realist view of cross language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language speech research* (pp. 171–204). York Press.
- Best, C., & Tyler, M. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In O. Bohn & M. J. Munro (Eds.), *Language experience in second language speech learning: In honor of James Emil Flege* (pp. 13–34). John Benjamins.

- Bilodeau, I. M. (1966). Information feedback. In E. A. Bilodeau (Ed.), *The acquisition of skill* (pp. 255–289). Academic Press.
- Bliss, H., Abel, J., & Gick, B. (2018). Computer-assisted visual articulation feedback in L2 pronunciation instruction: A review. *Journal of Second Language Pronunciation*, 4(1), 129–153.
<https://doi.org/10.1075/jslp.00006.bli>
- Celce-Murcia, M., Brinton, D. M., & Goodwin, J. M. (1996). *Teaching pronunciation: A reference for teachers of English to speakers of other languages*. Cambridge University Press.
- Cheng, C. -H., & Chen, C. -H. (2022). Investigating the impacts of using a mobile interactive English learning system on the learning achievements and learning perceptions of student with different backgrounds. *Computer Assisted Language Learning*, 35(1–2), 88–102.
<https://doi.org/10.1080/09588221.2019.1671460>
- Chen, X. (2017). 中国西班牙语专业学生语音层面语用失误及原因分析 [Analysis of pragmatic failure and causes in phonetic aspect among Spanish majors in China]. *广东外语外贸大学学报 [Journal of Guangdong University of Foreign Studies]*, 28(2), 54–62.
<https://gdwy.cbpt.cnki.net/WKH/WebPublication/wkTextContent.aspx?navigationContentID=37ef7ada-04c5-4e7d-ac0f-145f97269889&mid=gdwy>
- Chen, X., Li, J., & Zhu, S. (2021). Translanguaging multimodal pedagogy in French pronunciation instruction: Vis-a-vis students' spontaneous translanguaging. *System*, 101, Article 102603.
<https://doi.org/10.1016/j.system.2021.102603>
- Chen, Z. (2011). Errores articulatorios de los estudiantes chinos en la pronunciación de las consonantes españolas [Articulatory errors of Chinese college students on the pronunciation of Spanish consonants]. *Revista de Enseñanza de ELE a Hablante de Chino [Journal of Teaching Spanish as a Foreign Language to Chinese Native Speakers]*, 11(4), 55–67.
<https://www.sinoele.org/index.php/numeros/sinoele-4?id=137&lang=es>
- Chimitdorzhieva, G. N. (2018). Корневая морфема bur-/pur- в монгольских языках [The root morpheme bur-/pur- in the Mongolian languages]. *Sibirskii Filologicheskii Zhurnal [Siberian Journal of Philology]*, 3, 232–245. <https://doi.org/10.17223/18137083/64/21>
- Chun, D. M., Jiang, Y., Meyr, J., & Yang, R. (2015). Acquisition of L2 Mandarin Chinese tones with learner-created tone visualizations. *Journal of Second Language Pronunciation*, 1(1), 84–114.
<https://doi.org/10.1075/jslp.1.1.04chu>
- Cleland, J., Scobbie, J. M., & Wrench, A. A. (2015). Using ultrasound visual biofeedback to treat persistent primary speech sound disorders. *Clinical Linguistics & Phonetics*, 29(8–10), 575–597.
<https://doi.org/10.3109/02699206.2015.1016188>
- Dai, Y., & Wu, Z. (2021). Mobile-assisted pronunciation learning with feedback from peers and/or automatic speech recognition: A mixed-methods study. *Computer Assisted Language Learning*, 1–24.
<https://doi.org/10.1080/09588221.2021.1952272>
- de Bot, C. L. J. (1980). The role of feedback and feedforward in the teaching of pronunciation—an overview. *System*, 8(1), 35–45. [https://doi.org/10.1016/0346-251X\(80\)90022-6](https://doi.org/10.1016/0346-251X(80)90022-6)
- DeKeyser, R. (1997). Beyond explicit rule learning: Automatizing second language morphosyntax. *Studies in Second Language Acquisition*, 19(2), 195–221.
<https://doi.org/10.1017/S0272263197002040>
- DeKeyser, R. (2007). Study abroad as foreign language practice. In R. M. DeKeyser (Ed.), *Practice in a second language: Perspectives from applied linguistics and cognitive psychology* (pp. 208–226). Cambridge University Press.

- DeKeyser, R. (2017). Knowledge and skill in ISLA. In S. Loewen & M. Sato (Eds.), *The Routledge handbook of instructed second language acquisition* (pp. 15–32). Routledge.
- del Puy Ciriza, M., & Rivera-Campos, A. (2020). Teaching the Spanish trill to L1 English speakers using ultrasound instruction: A preliminary study on pronunciation pedagogy. *Journal of Spanish Language Teaching*, 7(1), 20–33. <https://doi.org/10.1080/23247797.2020.1770464>
- Derwing, T. M., & Munro, M. J. (2005). Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly*, 39(3), 379–397. <https://doi.org/10.2307/3588486>
- Deterding, D. (2007). *Singapore English*. Edinburgh University Press. <https://doi.org/10.3366/edinburgh/9780748625444.001.0001>
- Fernández, A. I. (2015). La corrección de la pronunciación de los estudiantes sinohablantes en el aula de E/LE[The correction of native Chinese students' pronunciation on Spanish as a foreign language] . *Foro de profesores de E/LE[Forum of Teachers of Spanish as a Foreign Language]*, 11, 189–196. <https://ojs.uv.es/index.php/foroele/index>
- Flege, J. E. (1995). Second language speech learning: Theory, findings and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 229–273). York Press.
- Flege, J., Aoyama, K., & Bohn, O. (2021). The revised speech learning model (SLM-r). In R. Wayland (Ed.), *Second language speech learning: Theoretical and empirical progress* (pp. 84–118). Cambridge University Press. <https://doi.org/10.1017/9781108886901.003>
- Foote, J. A., Trofimovich, P., Collins, L., & Soler Urzúa, F. (2016). Pronunciation teaching practices in communicative ESL classes. *The Language Learning Journal*, 44(2), 181–196. <http://dx.doi.org/10.1080/09571736.2013.784345>
- Fouz-González, J. (2020). Using apps for pronunciation training: An empirical evaluation of the English File Pronunciation app. *Language Learning & Technology*, 24(1), 62–85. <https://doi.org/10125/44709>
- Glaser, M., Lengyel, D., Toulouse, C., & Schwan, S. (2017). Designing computer-based learning contents: influence of digital zoom on attention. *Educational Technology Research and Development*, 65(5), 1135–1151. <https://doi.org/10.1007/s11423-016-9495-9>
- Guenther, F. H., Ghosh, S. S., & Tourville, J. A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language*, 96(3), 280–301. <https://doi.org/10.1016/j.bandl.2005.06.001>
- Haldin, C., Acher, A., Kauffmann, L., Hueber, T., Cousin, E., Badin, P., Perrier, P., Fabre, D., Perennou, D., Detante, O., Jaillard, A., Loevenbruck, H., & Baciú, M. (2018). Speech recovery and language plasticity can be facilitated by Sensori-motor fusion training in chronic non-fluent aphasia. A case report study. *Clinical Linguistics & Phonetics*, 32(7), 595–621. <https://doi.org/10.1080/02699206.2017.1402090>
- Hazan, V., Sennema, A., Iba, M., & Faulkner, A. (2005). Effect of audiovisual perceptual training on the perception and production of consonants by Japanese learners of English. *Speech Communication*, 47(3), 360–378. <https://doi.org/10.1016/j.specom.2005.04.007>
- Heift, T., & Chapelle, C. A. (2012). Language learning through technology. In S. M. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 555–569). Routledge.
- Kennedy, S., & Trofimovich, P. (2017). Pronunciation acquisition. In S. Loewen & M. Sato (Ed.), *The Routledge handbook of instructed second language acquisition* (pp. 260–279). Routledge.

- Khoroshilova, S. (2016). Anxiety in a foreign language pronunciation class in a university setting. In International Multidisciplinary Scientific GeoConferences (Ed.), *Proceedings of the 3rd international multidisciplinary scientific conference on social sciences and arts (SGEM 2016)* (pp.541–548). Curran Associates, Inc.
- Koen, M., Bitzer, E. M., & Beets, P. A. D. (2012). Feedback or feedforward? A case study in one higher education classroom. *Journal of Social Sciences*, 32(2), 231–242.
<https://doi.org/10.1080/09718923.2012.11893068>
- Kröger, B. J., Birkholz, P., Hoffmann, R., & Meng, H. (2010). Audiovisual tools for phonetic and articulatory visualization in computer-aided pronunciation training. In A. Esposito, N. Campbell, C. Vogel, A. Hussain., & A. Nijholt (Eds.), *Development of multimodal interfaces: Active listening and synchrony* (pp. 337–345). Springer. https://doi.org/10.1007/978-3-642-12397-9_29
- Levis, J., & Pickering, L. (2004). Teaching intonation in discourse using speech visualization technology. *System*, 32(4), 505–524. <https://doi.org/10.1016/j.system.2004.09.009>
- Li, S. (2008). 维吾尔语独特语音现象成因分析[Analysis of contributing factors in Uygurian special pronunciation]. *语言与翻译[Language and Translation]*, 2, 3–10.
- Li, Y. (1984). 鄂豫方言中的颤音 [The trill sound in Hubei and Henan provinces]. *华中师范大学学报 [Journal of Central China Normal University]*, 5, 121–125.
<http://journal.cnu.edu.cn/sk/CN/volumn/home.shtml>
- Li, Y., & Somlak, T. (2017). The effects of articulatory gestures on L2 pronunciation learning: A classroom-based study. *Language Teaching Research*, 23(3), 352–371.
<https://doi.org/10.1177/1362168817730420>
- Lieberman, A., & Whalen, D. (2000). On the relation of speech to language. *Trends in Cognitive Sciences*, 4(5), 187–196. [https://doi.org/10.1016/S1364-6613\(00\)01471-6](https://doi.org/10.1016/S1364-6613(00)01471-6)
- Lipetz, H. M., & Bernhardt, B. M. (2013). A multi-modal approach to intervention for one adolescent's frontal lisp. *Clinical Linguistics & Phonetics*, 27(1), 1–17.
<https://doi.org/10.3109/02699206.2012.734366>
- Liu, M. (2006). Anxiety in Chinese EFL students at different proficiency levels. *System*, 34(3), 301–316.
<https://doi.org/10.1016/j.system.2006.04.004>
- Liu, Y.-T., & Tseng, W.-T. (2019). Optimal implementation setting for computerized visualization cues in assisting L2 intonation production. *System*, 87, Article 102145.
<https://doi.org/10.1016/j.system.2019.102145>
- Long, A. Y. (2017). Investigating the relationship between instructor research training and pronunciation-related instruction and oral corrective feedback. In L. Gurzynski-Weiss (Ed.), *Expanding individual difference research in the interaction approach: Investigating learners, instructors, and other Interlocutors* (pp. 201–224). John Benjamins.
- Lu, J. (1991). 汉语和西班牙语语音对比—兼析各自作为外语学习的语音难点 [A comparison of speech sounds between Chinese Mandarin and Spanish and analysis of their respective difficulties in foreign language pronunciation learning]. *外国语 [Journal of Foreign Languages]*, 6, 58–63.
<http://jfl.shisu.edu.cn/CN/1004-5139/home.shtml>
- Lyster, R., & Sato, M. (2013). Skill acquisition theory and the role of practice in L2 development. In M. del Pilar García, M. J. G. Mangado, & M. Martínez-Adrián (Eds.), *Contemporary approaches to second language acquisition* (pp. 71–91). John Benjamins.

- Massaro, D. W., & Light, J. (2003). Read my tongue movements: bimodal learning to perceive and produce non-native speech /r/ and /l/. In *Proceedings of Eurospeech (Interspeech), 8th European conference on speech communication and technology* (pp. 1–4). Geneva, Switzerland.
- Mayer, R. E. (2009). *Multimedia learning* (1st ed.). Cambridge University Press.
<https://doi.org/10.1017/cbo9780511811678>
- Mayer, R. E. (2014). *The Cambridge handbook of multimedia learning* (2nd ed.). Cambridge University Press. <https://doi.org/10.1017/cbo9781139547369>
- Offerman, H. M., & Olson, D. J. (2016). Visual feedback and second language segmental production: The generalizability of pronunciation gains. *System*, 59, 45–60.
<https://doi.org/10.1016/j.system.2016.03.003>
- Patten, I., & Edmonds, L. A. (2013). Effect of training Japanese L1 speakers in the production of American English /r/ using spectrographic visual feedback. *Computer Assisted Language Learning*, 28(3), 241–259. <https://doi.org/10.1080/09588221.2013.839570>
- Pérez-Vidal, C. (2017). Study abroad and ISLA. In S. Loewen & M. Sato (Eds.), *The Routledge handbook of instructed second language acquisition* (pp. 339–360). Routledge.
- Rogerson-Revell, P. M. (2021). Computer-assisted pronunciation training (CAPT): current issues and future directions. *RELC Journal*, 52(1), 189–205. <https://doi.org/10.1177/0033688220977406>
- Saito, K., & Lyster, R. (2012). Effects of form-focused instruction and corrective feedback on L2 pronunciation development of /l/ by Japanese learners of English. *Language Learning*, 62(2), 595–633. <https://doi.org/10.1111/j.1467-9922.2011.00639.x>
- Saito, K., & Plonsky, L. (2019). Effects of second language pronunciation teaching revisited: A proposed framework and meta-analysis. *Language Learning*, 69(3), 652–708.
<https://doi.org/10.1111/lang.12345>
- Sato, M., & Lyster, R. (2012). Peer interaction and corrective feedback for accuracy and fluency development: Monitoring, practice, and proceduralization. *Studies in Second Language Acquisition*, 34(4), 591–626. <https://doi.org/10.1017/S0272263112000356>
- Schmidt, R. (1993). Consciousness, learning and interlanguage pragmatics. In G. Kasper & S. Blum-Kulka (Eds.), *Interlanguage pragmatics* (pp. 43–57). Oxford University Press.
- Seibert Hanson, A. E., & Brown, C. M. (2019). Enhancing L2 learning through a mobile assisted spaced-repetition tool: An effective but bitter pill? *Computer Assisted Language Learning*, 33(1–2), 133–155. <https://doi.org/10.1080/09588221.2018.1552975>
- Sung, H. -Y., Hwang, G. -J., Lin, C. -J., & Hong, T. -W. (2017). Experiencing the Analects of Confucius: An experiential game-based learning approach to promoting students' motivation and conception of learning. *Computers & Education*, 110(C), 143–153. <https://doi.org/10.1016/j.compedu.2017.03.014>
- Suzuki, Y., Nakata, T., & DeKeyser, R. M. (2019). The desirable difficulty framework as a theoretical foundation for optimizing and researching second language practice. *Modern Language Journal*, 103(3), 713–720. <https://doi.org/10.1111/modl.12585>
- Swain, M. (1995). Three functions of output in second language learning. In G. Cook & B. Seidlhofer (Eds.), *Principles and practice in applied linguistics* (pp. 125–144). Oxford University Press.
- Toutouchi, F., & Izquierdo, E. (2018). Enhancing digital zoom in mobile phone cameras by low complexity super-resolution. In *Proceedings of the 2018 IEEE international conference on multimedia & expo workshops (ICMEW)* (pp. 1–6). IEEE.
<https://doi.org/10.1109/ICMEW.2018.8551540>

- van Gog, T. (2014). The signaling (or cueing) principle in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 263–278). Cambridge University Press.
<https://doi.org/10.1017/CBO9781139547369.014>
- VanPatten, B., & Williams, J. (2015). Early theories in SLA. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (2nd ed., pp. 11–33). Routledge.
- Walley, A. C., Flege, J. E., & Randazza, L. A. (1993). Effects of lexical status on children's and adults' perception of native and non-native vowels. *The Journal of the Acoustical Society of America*, 93(4_Supplement), 2297. <https://doi.org/10.1121/1.406508>
- Wrembel, M. (2013). Foreign accent ratings in third language acquisition: The case of L3 French. In E. Waniek-Klimczak & L. R. Shockey (Eds.), *Teaching and researching English accents in native and non-native speakers* (pp. 31–47). Springer. https://doi.org/10.1007/978-3-642-24019-5_3
- Wrembel, M. (2007). "Still sounds like a rainbow" - a proposal for a coloured vowel chart. In *Proceedings of the phonetics teaching and learning conference PTLC2007* (pp. 1–4). University College London.
- Wu, M.-H. (2021). The applications and effects of learning English through augmented reality: a case study of *Pokémon Go*. *Computer Assisted Language Learning*, 34(5-6), 778–812.
<https://doi.org/10.1080/09588221.2019.1642211>
- Zhu, J., Zhang, X., & Li, J. (2022). Using AR filters in L2 pronunciation training: practice, perfection, and willingness to share. *Computer Assisted Language Learning*. Advance online publication.
<https://doi.org/10.1080/09588221.2022.2080716>

Appendix A. Stimuli for Production Test

Target segment	Stimuli			
/θ/	zapato	cine	zutano	zona
	zumo	cena	luce	mozo
	lozano	Pecio	usted	salud
	Mitad			
/ð/	pide	todo	sida	edad
	modo	nada	modelo	
/r/	ruta	rifa	Roma	remo
	ramo	resina	parra	torre
	farruto	terreno		
/l/	pulpo	olma	Nepal	pajil
	piel	mil	alma	olfato
	túnel	fatal		

Appendix B. Stimuli Used in Teaching

Target segment	Stimuli			
/θ/	ceba	circa	zoca	comed
	ece	ici	celoso	tomad
	piscina	aza	amenaza	cuidad
/ð/	zábida	aletazo	azul	azulado
	uzu	unidad	soledad	ozo
/r/	reno	rito	rosa	rutina
	rata	sonrisa	alrededor	las risas
	perro	narrativa	carreta	farruco
	enrase	Enrique		
/l/	ful	nivel	adelfa	nalga
	a mal	babel	terral	bulbo
	estatal	al día	del mismo	

Appendix C. Questions of Semi-structured Interview

Original version (Chinese)	Translated version (English)
<p>你认为语音课在学习发音上对你有帮助，请举例说明。</p>	<p>In what way do you think this pronunciation session helps your pronunciation? please give some examples.</p>
<p>你是否注意到该课程的授课方式有什么特别之处？你觉得这种方式怎么样？请具体谈一谈优点和缺点。</p>	<p>Did you notice something different in the way the instruction was given, how do you like it? Please give the advantages and disadvantages, if any.</p>
<p>你觉得这种方式在哪些方面帮助了你的发音？</p>	<p>In what way do you think this strategy facilitates your pronunciation?</p>
<p>你以后自己练发音的时候会使用这种方法吗？</p>	<p>Will you apply this strategy next time when you practice pronunciation by yourself? Why or why not?</p>
<p>你是否觉得上完课之后自己在发音时会格外注意发音位置和口型？请举例说明。</p>	<p>Did you pay extra attention to articulatory gestures in your pronunciation practice after the class? Please give examples.</p>
<p>具体在技术应用层面，你对语音课的组织形式有什么自己的想法吗？</p>	<p>Do you have any ideas or suggestions regarding technology application for pronunciation class?</p>

About the Authors

Siqi Wang is a graduate student in applied linguistics at Shanghai University of Finance and Economics. Her research interests include phonetics, phonology, and the use of mobile technologies in L2 pronunciation learning and teaching.

E-mail: lesleybroccoli47@gmail.com

Jian Li (corresponding author) is an associate professor at School of Foreign Studies, Shanghai University of Finance and Economics. She conducts research on the effects of mobile-assisted technology (e.g., digital zoom, games, AR, online videoconference) on L2 pronunciation learning and teaching. Jian Li is the corresponding author.

E-mail: li.jian@mail.shufe.edu.cn

ORCID: 0000-0001-7544-5344

Qian Liang is a Ph.D. student at Language Institute, Complutense University of Madrid, and a lecturer of Spanish at Shanghai University of Finance and Economics. Her research focuses on teaching Spanish as a foreign language and metaphorical cognition.

E-mail: dorschgolden@gmail.com