Language Learning & Technology
ISSN 1094-3501 CC BY-NC-ND

October 2023, Volume 27, Issue 3
pp. 83–107

ARTICLE

Restructuring multimodal corrective feedback through Augmented Reality (AR)-enabled videoconferencing in L2 pronunciation teaching

Yiran Wen, Shanghai University of Finance and Economics

Jian Li, Shanghai University of Finance and Economics

Hongkang Xu, Northeastern University

Hanwen Hu, Shanghai University of Finance and Economics

Abstract

The problem of cognitive overload is particularly pertinent in multimedia L2 classroom corrective feedback (CF), which involves rich communicative tools to help the class to notice the mismatch between the target input and learners' pronunciation. Based on multimedia design principles, this study developed a new multimodal CF model through augmented reality (AR)-enabled videoconferencing to eliminate extraneous cognitive load and guide learners' attention to the essential material. Using a quasi-experimental design, this study aims to examine the effectiveness of this new CF model in improving Chinese L2 students' segmental production and identification of the targeted English consonants (dark /ɫ/, /ð/and /θ/), as well as their attitudes towards this application. Results indicated that the online multimodal CF environment equipped with AR annotation and filters played a significant role in improving the participants' production of the target segments. However, this advantage was not found in the auditory identification tests compared to the offline CF multimedia class. In addition, the learners reported that the new CF model helped to direct their attention to the articulatory gestures of the student being corrected, and enhance the class efficiency. Implications for computer-assisted pronunciation training and the construction of online/offline multimedia learning environments are also discussed.

Keywords: *Corrective Feedback, Multimedia Learning, Computer-assisted Pronunciation Training (CAPT), AR-enabled Videoconferencing*

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

APA Citation: Wen, Y., Li, J., Xu, H., & Hu, H. (2023). Restructuring multimodal corrective feedback through Augmented Reality (AR)-enabled videoconferencing in L2 pronunciation teaching. *Language Learning & Technology*, *27*(3), 83–107. https://hdl.handle.net/10125/73533

Introduction

New emerging technologies, such as computer-assisted language learning (CALL) or artificial intelligence (AI), have recently been researched and adopted to provide strong support for the overall effectiveness of corrective feedback (CF), which plays an important role in second language (L2) teaching (Klimova & Pikhart, 2022). Despite the growing interest in the exploration of how technologies may contribute to providing digital or automatic CF, almost no research study has seriously examined how such technologies can facilitate the process of giving explicit CF to students' pronunciation output particularly in L2 classroom settings. By providing CF in teaching pronunciation, teachers can help improve L2 learners' English pronunciation and encourage them to reflect on their strengths and weaknesses in pronunciation (Agustuna et al., 2019). In comparison with listening only to model pronunciation and the students' own recordings, intelligibility and comprehensibility have been shown to improve when individual CF is presented, resulting in higher accuracy of pronunciation (Dlaska & Krekeler, 2013; Saito, 2015). However, if the individual CF cannot be shared to the rest of the class both auditorily and visually, the class efficiency will be greatly harmed. Thus, methods to engage the rest of the class when providing individual CF needs to be addressed in research on teaching L2 pronunciation.

Thanks to the rapid development of technology, multimedia learning environments are enriched and have so far included “online instructional presentations, interactive lessons, e-courses, simulation games, virtual reality, and computer-supported in-class presentations” (Mayer, 2014, p. 281), enabling more multimodality in the classroom. However, though multimedia environments may boost language comprehension and interaction, a new challenge for multimedia learning research grounded in cognitive theory was raised by Mayer and Fiorella (2014). They noted that due to the limit of learners’ cognitive capacity, multimedia learning environments are sometimes so overwhelming that the learner has to process a lot of extraneous material and may not be able to allot the required cognitive time and effort to processing essential material. This problem, namely cognitive overload, is particularly pertinent in multimedia classrooms of L2 pronunciation CF since learners are required to constantly compare their own/peers’ output with the target language input, receive elaboration from the teacher, and communicate with the teacher at the same time during the classroom CF. There is thus a need for further investigation to optimize multimodal tools in order to alleviate learners’ cognitive load in multimedia environments.

This study aims to develop a new model through augmented reality (AR)-enabled videoconferencing to facilitate CF in L2 pronunciation classes. Based on multimedia design principles, the model integrates different types of computer tools including videoconferencing, AR annotation and filters to eliminate extraneous cognitive load, guide learners’ attention to the essential material and maximize classroom CF efficiency. The present study investigates whether this new multimodal CF model of computer-assisted pronunciation training (CAPT) is beneficial for L2 learners in improving their segmental production and perception.

Literature Review

CF Efficiency Problems in Traditional Multimedia Pronunciation Classrooms

CF, defined as teachers’ responses to learners’ utterances when their use of the target languages is incorrect, is a pivotal language teaching tool in the L2 classroom, boosting feedback and interaction between teachers and students (Heift & Nguyen, 2021; Lyster et al., 2013). CF provides support in drawing the learner’s attention to discrepancies between their output and the target language input, and is therefore assumed to improve their knowledge of the target language feature and facilitate second language acquisition (SLA) (Mackey, 2006).

However, previous studies have revealed three unresolved pedagogical problems in using multimedia for L2 pronunciation teaching and CF, relating to cognitive overload and affecting correction efficiency. First, the physical seating arrangements in traditional language classrooms mean that students seated at the back are at a disadvantage in terms of class participation, since their visual contact with both the teacher and the learner being corrected are relatively obstructed (Sommer, 1967). This visual deficiency caused by physical classroom space creates difficulties not only for teachers in attracting the observers’ attention and maintaining their participation in CF, but also for students acting as listeners/observers in integrating the auditory pronunciation output of the learner being corrected with his/her visual articulations. This clearly reduces the maximal effectiveness of classroom CF, since some studies have found that classmates can also learn from corrections that address another student’s incorrect utterance in the classroom context (Havranek, 2002; Kim & Han, 2007).

Second, as Navarra and Soto-Faraco (2007, p. 4) note, “speech is a multimodal phenomenon in which the articulatory movements of the speaker produce correlated information in vision (i.e., lip movements) and audition (linguistic sounds)”. The early studies have shown that L2 listeners often fail to hear the difference between certain non-native phonemic contrasts (Best, 1995a; Flege, 1995). However, traditional multimedia language labs are equipped with language learning headsets that block outside sounds and disturbances, a media player for listening to and watching audio and visual materials, and recording tools for recording pronunciation, constituting a system which is heavily weighted in favor of audio support. Visual reinforcement tools are needed to help with the selection of visual images of articulatory gestures that teachers need in order to detect learners' articulatory errors, and students need to perceive mismatches between the target input and their own pronunciation (Mackey, 2006). Also, learners need a concurrent on-screen demonstration of the incorrect articulation being corrected, and pre-recorded native pronunciation input to make the most of their limited time and effort devoted to noticing the gap (Mayer & Fiorella, 2014).

Third, although most L2 learners understand that correction is essential and most teachers assume that in-class CF is important for both observers and the learner being corrected (Havranek, 2002; Schulz, 2001), negative affective factors such as language anxiety and embarrassment may outweigh the positive effects of CF (Kartchava & Ammar, 2013; Lee, 2016). This is particularly true in pronunciation and phonetics classes, where learners are often asked to perform and be directly assessed in front of others (Baran-Lucarz, 2013; Khoroshilova, 2016). As such, the traditional multimedia language classroom badly needs more effective tools to weed out negative material that is irrelevant to articulatory and phonetic knowledge.

The above-mentioned three issues present in the traditional offline L2 pronunciation classrooms have led to the teacher's limited use of individual CF in order to guarantee class participation, and these issues also provide the context in which we developed our study. To establish the empirical foundation of our study, we review the literature on CAPT and articulatory-based visual aids used for CF within the framework of L2 pronunciation acquisition as follows.

CAPT and Articulatory-based Visual Feedback

New technology, new material and new media have constantly promoted the development of CAPT, which has extended beyond traditional language lab to mobile devices, automatic speech recognition (ASR), speech visualization, artificial intelligence (AI), automatically providing assessment and visual feedback for learners to improve their pronunciation (Engwall, 2012; Rogerson-Revell, 2021; Tsai, 2019). While the bulk of the work in CAPT research has examined how to train learners' connected speech by displaying visual representations of pitch/intonation contours (Chun et al., 2015; Levis & Pickering, 2004), waveforms, and spectrograms (Liu & Tseng, 2019; Patten & Edmonds, 2015), few studies have investigated the utility of visual feedback on segmental production in classroom settings (Olson, 2014). Moreover, as Rogerson-Revell pointed out, many CAPT systems provide visual displays of raw data that require "some degree of expertise or training to be interpreted" (2021, p. 193), leaving the learners puzzled without the guidance of a teacher. Also, this type of acoustic-based visual feedback may give the learners automated scoring but fail to tell them the causes of their errors and how to correct their production. Therefore, there is an urgent need to develop a CAPT system incorporating audiovisual feedback to illustrate the articulatory gestures of the sound so that the learner can rely on this immediate, individualized and accessible visual feedback to correct their pronunciation errors.

The multimedia design of this study is built on an established motor theory of speech perception, which makes a strong claim that visual (gestural) speech information can impact the perception of spoken language (Liberman et al., 1967; Liberman & Mattingly, 1985; Liberman & Whalen, 2000). Meanwhile, the fuzzy logical model of perception (FLMP) concludes that a combination of multisensory cues may enhance perception, with better effects than each sensory modality in isolation (Massaro, 1998). Previous psychological studies have revealed that visually specific gestures are usually integrated with auditory sounds and may facilitate spoken speech perception (van Wassenhove et al., 2005).

It remains unresolved whether the "articulatory-based audiovisual integration of speech can facilitate the perception of phonological contrasts in a second language" (Navarra & Soto-Faraco, 2007, p. 5), but according to the perception assimilation model (PAM) (Best, 1994; 1995a; 1995b) and PAM-L2 (Best & Tyler, 2007), L2 learners can only achieve an accurate perception of L2 sounds through the identification of the articulatory gestures involved in the production of these sounds. In Fouz-González's (2015) review of CAPT research, animated talking heads that illustrated articulatory knowledge were accounted as a good example to prove the effect of articulatory-based audiovisual information on learners' pronunciation enhancement both in perception and production. Engwall (2008) reported positive effects in articulatory instructions given by a talking head to illustrate the L2 pronunciation of two Swedish phonemes /r/ and /ɧ/. Li and Somlak (2017) compared the effects of audio-only and audio-visual input in L2 pronunciation teaching, and found that students in the multimedia environment with audio-visual aids performed significantly better in the pronunciation of the target contrasts. Therefore, although cognitive overload may be an issue in CF classes with multimedia, this study does not propose a return to the audio-only teaching tradition (e.g., the listen-and-repeat mode). Instead, we adopt AR technology, which may be integrated into a videoconferencing platform to facilitate the selection, organization, and integration of the articulatory-based audiovisual information that is essential for the correction of target sounds.

Reconstructing Online AR Multimedia Learning Model

Cognitive Principles of Multimedia Learning

In the light of the aforementioned research gaps in articulatory-based visual aids facilitating individual CF in L2 pronunciation teaching, this study draws on cognitive principles in multimedia learning in the reconstruction of an online CF system via videoconferencing and AR technology in order to resolve the cognitive load issues and the conflict between individual CF and classroom efficacy. The four cognitive principles of multimedia design --the coherence principle, the signaling principle, the spatial contiguity principle, and the temporal contiguity principle (de Koning et al., 2009; Mayer & Moreno, 2003; Mayer & Fiorella, 2014; van Gog, 2014) are defined as follows:

1) The coherence principle refers to load-reducing techniques eliminating interesting but extraneous material.

2) The signaling principle, also known as the cueing principle, refers to learning reinforcement when multimedia messages are incorporated with cues and signals that guide attention to the essential material.

3) The spatial contiguity principle is that "people learn more deeply from a multimedia message when corresponding words and pictures are presented near rather than far from each other" (Mayer & Fiorella, 2014, p.303), in order to reduce the need for visual scanning.

4) The temporal contiguity principle refers to learning reinforcement when multimedia messages (visual and verbal) are presented synchronously rather than successively.

Videoconferencing Technology and AR Technology

The history of videoconferencing can be traced back to the twentieth century, but it became particularly prevalent in the field of higher education in the last two decades (Correia et al., 2020). It offers a synchronous channel that enables immediate interaction between teachers and students, facilitating instant feedback and supporting information transformation between people (Wiesemes & Wang, 2010). The multimedia capacities of videoconferencing involve information presented in different modes such as audio, visual, and verbal, increasing students' overall learning efficacy (Correia et al., 2020). In the field of language teaching, Wang (2004) has argued that paralinguistic cues such as body language, facial expressions, and head nods can be caught during videoconferencing, reducing ambiguity in speech and enhancing understanding.

AR refers to technology that composites virtual elements in a real-world environment for the purpose of supplementing reality (Azuma, 1997; Carmigniani & Furht, 2011; Yeh & Tseng, 2020). In the recent years, we have witnessed the emergence and wide application of AR in education. Wu et al. (2013) summarized two characteristics of AR for educational purposes: first, AR helps learners to engage in real-world learning environments by providing virtual objects to help them to learn more deeply; second, AR integrates more digital resources with the real world than the traditional multimedia teaching environment. These educational values are closely related to how AR is designed, applied, and implemented in the multimedia learning context. Using AR annotation, information can be attached to specific objects to achieve augmentation.

As well as overlaying digital objects on a real environment, AR, such as the application of AR filter, can also function by removing or hiding real objects (Azuma, 1997). As yet, AR technology has rarely been used in pronunciation teaching, not to mention CF in a pronunciation class. Only Zhu et al. (2022) investigated the effectiveness of using an AR filter app during pronunciation practice and found that the app played a positive role in promoting students' segmental production and raising their articulatory awareness. Also, this AR filter app may not only draw students' attention to essential articulators, but also reduce students' shyness and anxiety when interacting with teachers.

In Table 1, the descriptions of the effects of the three technologies (videoconferencing, AR filters, and AR annotation) are summarized based on the cognitive theory of multimedia learning and their potential in addressing classroom CF issues in L2 pronunciation teaching. First, videoconferencing shortens the spatial distance between learning materials and the teacher's standard articulation model by displaying them on a single screen, reducing students' visual scanning. In addition, a videoconferencing platform enables synchronous presentation of the teacher's narration and a model's animation. Taking advantage of the temporary contiguity principle, this reduces representational holding during essential processing. Second, AR filters with masks can be designed based on the coherence principle, reducing the cognitive burden caused by non-relevant facial expressions. Also, students' embarrassment and unwillingness to show their articulators in front of the class may be reduced (Zhu et al., 2022). Additionally, the digital zoom technology integrated in AR filters may enlarge the articulators, guiding both the teacher's and students' attention to the relevant elements and thereby helping them to detect articulatory errors. Finally, according to the signaling principle, AR annotation of different points of articulation will cue students' attention to key elements.

Table 1

Descriptions of the Technological Affordances on Multimedia Design Principles

Technologies adopted	Multimedia principles	Descriptions of the technological affordances
Videoconferencing	Contiguity principle	Spatial contiguity effect: better transfer when learning materials are close together. Temporal contiguity effect: the platform enables synchronous narration and animation
AR filter	Coherence principle	Extraneous unwanted facial expressions are excluded.
	Signaling principle	Essential articulation materials are enlarged and reinforced.
AR annotation	Signaling principle	Signals may be included on crucial articulators to facilitate noticing.

The Present Study

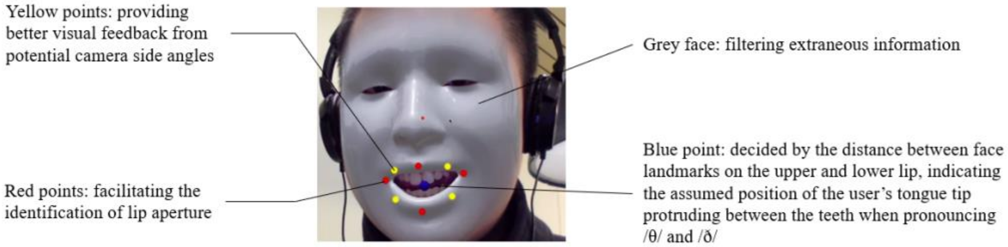
Design

Based on the multimedia principles discussed above, we develop a model to eliminate extraneous cognitive load caused by the multimedia context. This study adopts the Tencent Meeting app (TMA), the most commonly-used videoconferencing app in China. This app has also been adopted by many universities because of its high-quality HD video and the other functions it offers (Talidong, 2020).

Although the app does not provide the camera that we need to reinforce articulatory gestures and reduce the cognitive load in CF, TMA does enable users to install plug-in components. This study relies on a custom tool, since none of the freely available tools were able to integrate both AR annotation and AR filters into one camera lens. Snap Camera provides a plug-in virtual camera that enabled us to apply any filter. Using Lens Builder, a tool made by Snap AR and Snapchat, we could create, publish, and share tailored AR filters which could then be utilized via Snap Camera and displayed real-time on TMA. To customize the AR annotation lens focusing on the articulators, we used the "Face Landmarks" database offering 93 tracked points on the user's face and selected eight of these anchored-to-static points spanning the oral commissures, the Cupid's bow, and various other middle points along the vermilion border (see Figure 1).

Figure 1

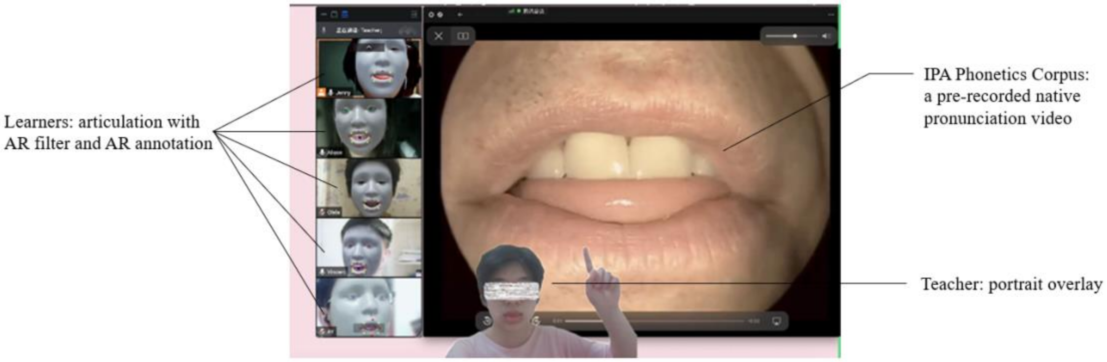
The Customized Lens with AR Annotation and Filter


Yellow points: providing better visual feedback from potential camera side angles
Grey face: filtering extraneous information
Blue point: decided by the distance between face landmarks on the upper and lower lip, indicating the assumed position of the user's tongue tip protruding between the teeth when pronouncing /θ/ and /ð/
Red points: facilitating the identification of lip aperture

This custom tool was then plugged into TMA, and learners could see their images with the articulators-annotated AR filter on the screen. During the correction process the teacher used the computer screen to share a video articulation of the target segment by native speakers selected from the IPA Phonetics corpus. The corpus, as an asynchronous learning resource, is used as a reference to standard articulation models, in alignment with the students' synchronous articulations. With the help of this alignment on the screen and the annotated and enlarged articulators, the teacher delivered individualized CF by identifying differences between the model and the student's output. Other classmates could observe the whole CF process on TMA (see Figure 2).

Figure 2

Screenshot of the Online CF Class Through AR-enabled Videoconferencing


Learners: articulation with AR filter and AR annotation
IPA Phonetics Corpus: a pre-recorded native pronunciation video
Teacher: portrait overlay

Target Segments

Studies in laboratory settings have found that training with audio-visual aids may only benefit pronunciation when the articulatory features of the target sounds are sufficiently salient visually (Hazan et al., 2005). For example, "the visual correlate of the gesture associated with the pronunciation of /p/ is more salient than that associated with /k/" (Navarra & Soto-Faraco, 2007, p. 5). This is also supported by van Wassenhove et al.'s (2005) study through the lens of neural processing, which found that the saliency of visual input may impact the potential of the visual code in facilitating speech perception. Badin et al. (2010) not only found a positive effect of visual displays on segmental production, but noticed that a frontal view of the face was better perceived than a cutaway view of the head. This study selects the English phonemes /θ/, /ð/, and dark /ɫ/ as the target sounds, since these consonants are not only found to be difficult for Mandarin speakers to pronounce (Deterding 2006; 2007), but also have relatively visually distinguishable articulatory gestures (Hardison, 2003), especially involving a frontal view of the lips and the tongue. There are no dental consonants in the Mandarin phonetic inventory, so Mandarin speakers have difficulty placing the tip of the tongue between the top and bottom front teeth and thus usually pronounce /θ/ and /ð/ into the articulatorily-similar alveolar /s/ and /d/. Dark /ɫ/, which requires the tongue tip to reach up and close to the bony ridge and the lips to relax, is often mispronounced by Mandarin speakers as back vowel /ʊ/ when following a segment other than /ʊ/ (e.g., fill [fɪʊ], apple [æpʊ]) (Deterding, 2007). As we observed, in the correction of their erroneous place of articulation, Mandarin speakers tend to uncontrollably round their lips in the same way as they pronounce /ʊ/ and often fail to lift their tongue tip up. Therefore, we hypothesize that the annotation function of AR will help the students notice the discrepancies between the standard and their own articulatory gestures and better control their lips and tongues.

Research Questions

Our examination of the effectiveness of using AR-enabled videoconferencing in pronunciation CF will address the two sides of L2 acquisition (i.e., production and perception) as well as the learners' attitudes:

1) Does multimodal CF through AR-enabled videoconferencing improve the participants' production of the targeted consonants?

2) Does multimodal CF through AR-enabled videoconferencing facilitate the participants' identification of the targeted consonants?

3) What are the participants' attitudes towards the application of this new CF model?

Method

Participants

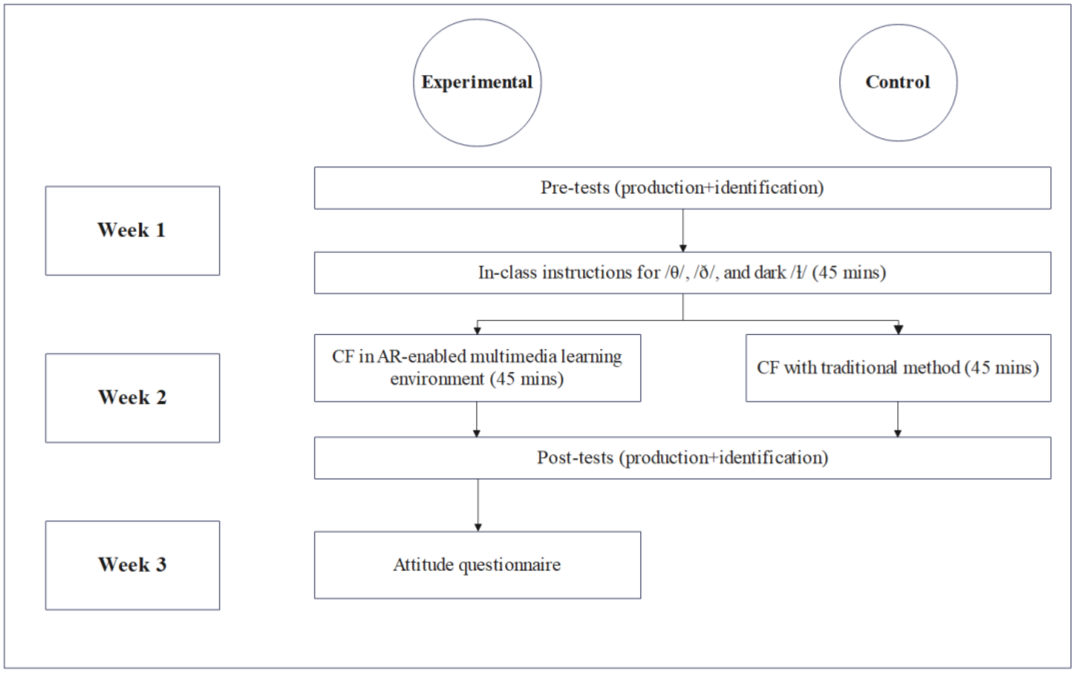
The participants were non-English major students enrolled in an English phonetics elective course during the first semester of their sophomore year at a first-tier university in Eastern China. They were 60 sophomores, 16 males and 44 females, from different majors including communication and media studies, law, sociology, accounting, and financial management. All reported normal hearing and vision. One natural class was randomly allocated to the experimental condition (N = 30), the other was allocated to the control condition (N = 30). Background questionnaires collected at the beginning of the study indicated that the students in the two groups were comparable overall in terms of their English proficiency, reporting an even distribution of English levels (A, B, C, D) in the English entrance examination.

Procedure and Devices

Both the experimental group and the control group were enrolled in a 16-week English phonetics course. The quasi-experiment lasted 3 weeks. In the first week both groups were asked to take pre-tests involving both production and identification tasks, and received 45 minutes of explicit instruction in the phonetic and articulatory knowledge related to the three target consonants /θ/, /ð/, and dark /ɫ/, using the same teaching method and the same traditional offline language teaching multimedia environment. In the second week both groups received 45 minutes of CF of different approaches, followed by intermediate post-tests. In the third week the experimental group were asked to fill out a questionnaire aimed at collecting opinions related to the new CF approach (see Figure 3).

Figure 3

Quasi-experimental Procedure


Experimental
Control
Week 1
Week 2
Week 3
Pre-tests (production+identification)
In-class instructions for /θ/, /ð/, and dark /ɫ/ (45 mins)
CF in AR-enabled multimedia learning environment (45 mins)
CF with traditional method (45 mins)
Post-tests (production+identification)
Attitude questionnaire

Both groups used IPA Phonetics corpus as a reference to the standard articulation model. The experimental group received CF from the teacher in an online classroom aided by videoconferencing, AR filters, and annotation technologies (see Figure 2). The experimental group was asked to plug the Snap camera with the AR filters and annotations into the videoconferencing technology and examine their devices one day before the experiment. During the class, each participant was required to switch on their AR-enabled virtual camera when it was his/her turn for one-on-one CF. The teacher and the students wore headphones or earphones at all times. In contrast, the control group received traditional CF in an offline multimedia classroom installed with Lancoo Language Learning Platform, where the standard animated model was shown on the screen of the desktop computer in front of every student, connected to and controlled by a central computer on the stage. The teacher and the students were also connected via their headphones. The teacher provided individualized CF to each student by listening to their auditory output with language learning headsets and detecting the articulation with their naked eyes. The tests for both groups were all administered in the same environment where the control group received CF. Their productions were audio-recorded using a Roland R-5 recorder in the studio, and also videotaped on their own digital devices. Videos were used as supporting articulatory evidence when raters had difficulty judging the accuracy by listening alone.

Evaluation

Production Tests

Both controlled and spontaneous tasks should be adopted in empirical studies to confirm whether pronunciation teaching leads to acquisition and changes in performance (Saito & Plonsky, 2019). Therefore, in our production test participants were asked to finish two tasks, a word-reading task, and an impromptu translation task. The first task involved reading aloud 12 words using the targeted sounds (see Appendix A) in order to test the participants' basic command of these sounds. The syllable position of the target sound in each word was also taken into consideration. Each target sound (/θ/, /ð/, dark / ɫ /) was presented in four words. The participants were not given any preparation time or allowed to repeat or correct their pronunciation in the word-reading task.

The impromptu translation task was used to elicit their subconscious pronunciation of the target segments in more natural speech. We used a semi-structured translation task with written Chinese prompts rather than a picture-description/naming task, an event-description task, or a question-answering task which are more extemporaneous and unstructured, since we were intended to precisely elicit the participants' use of the target segments. The planning time for the translation task was controlled to 5 seconds, after which the participants were requested to translate six Chinese sentences into English, testing each target consonant 5 times in total. Target words were chosen to make sure there were no alternative words for the same meaning. The sentences were taken from a beginners' English learning book, which guaranteed that the participants would be able to complete the task within the time limit while paying "simultaneous attention to the grammatical, phonological, lexical, and pragmatic aspects of language to convey their intended message" (Saito & Plonsky, 2019, p. 666).

Identification Tests

To test the participants on their accuracy in identifying the target segments in L1/L2 contrasts, forced-choice identification tests (see Appendix B) were adopted, which included two tasks. First, the participants were required to listen to 15 words (5 for each target sound), and identify whether there was any mistake in the pronunciation. If so, they were required to point out the mistake, write down the mispronounced segment, and correct it. Then the participants were asked to complete 9 multiple choice questions. There was only one correct answer to each question, with the other three choices serving as distracters. They heard 9 non-words containing the target sounds, which were either correctly pronounced or mispronounced. We used non-words to eliminate the interference from word meaning. The participants had to be able to distinguish the slight differences between the correct and incorrect pronunciations, which were common mistakes that Mandarin speakers might make in their daily communication, such as /θ/ mispronounced as /s/, /ð/ as /z/ or /d/, and /ɫ/ as /u/ or /o/ (Deterding, 2006; 2007). Considering the findings in high variability phonetic training (see Thomson, 2011 for a review of HVPT research) and Golden Speaker Builder (see Ding et. al., 2019 for a review of GSB training), we invited 2 native speakers to record the standard stimuli and 3 non-native speakers to record the mispronounced items, thus offering input from multiple voices.

Questionnaire

A questionnaire was administered to provide insights into the experimental group participants' opinions about the intervention (Pawlak et al., 2021). In this study we developed the questionnaire as a judging matrix to collect students' learning attitudes towards the AR-enabled CF model according to different dimensions, based on a scale that was specifically designed for the investigation of students' attitudes towards AR applications. The scale has been shown to have high validity and reliability (Küçük et al., 2014).

The questions were categorized into items related to instructional approach, learning quality, learning willingness, learning anxiety, learning difficulties, and learning interaction (see Appendix C). There were 3 questions for each category, including one negative/reverse-wording item and two positive items. Reverse-wording items were included to avoid possible acquiescence bias and check whether the respondents were giving consistent answers. A 5-point Likert scale was used to measure the participants' levels of agreement with the statements, from high to low with one neutral option in the middle. Additionally, the last question in the questionnaire was open-ended in order to elicit the students' candid opinions on their overall experience, and collect any suggestions they might have for the improvement of our teaching model.

Data Analysis

The participants' production was rated by two experienced EFL teachers (a native speaker and a Mandarin-English bilingual), both with doctoral degrees in Linguistics. The ratings were dichotomous: 1 if the target sound was pronounced correctly, 0 if it was mispronounced. The raters could listen to the recordings repeatedly and check the videotaped production if they were having difficulty in deciding. Interrater reliability was established at 0.91 using Cohen's kappa. To determine the intra-rater reliability the two raters remarked the pre-tests and post-tests of three of the participants two months later, yielding a reliability figure of 0.94.

To examine whether our results were suitable for ANOVA or other statistical methods, a normality test was necessary to determine the distribution of the scores. Kolmogorov-Smirnov tests showed that the results from both the production tests ($p = .069 > .05$) and the identification tests ($p = .197 > .05$) followed normal distributions. Thus, this study used two-way ANOVA, independent samples t-tests and paired t-tests to analyze the quantitative production and identification data. Following Martin (2020), independent samples t-tests were used to compare the pre-test scores between the experimental group and the control group. The results confirmed that the two groups were matched before the treatment: t (58) = -1.509, $p = .13 >.05$.

Quantitative and qualitative analyses were both adopted when dealing with the data from the questionnaires. Mean scores for each scale were calculated to obtain an overall understanding of the participants' attitudes. The questionnaires were analyzed using SPSS 25.0, including calculations of reliability, mean and standard deviation. Negative items were transformed via reverse-coding in SPSS.

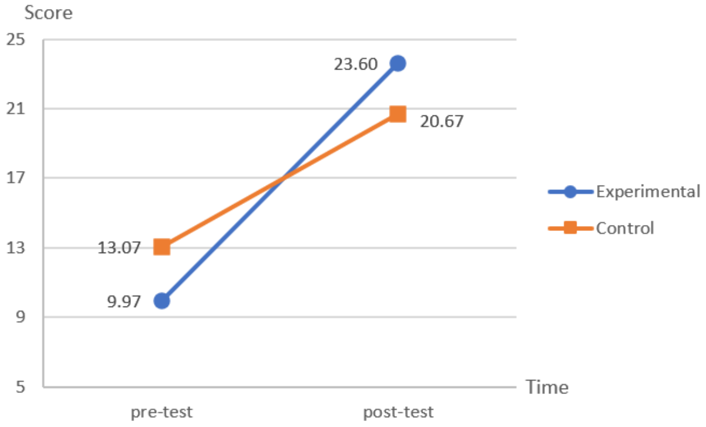
Results

Production Tests

Figure 4 graphically presents the mean changes for the experimental and control groups in the pre- and post-tests in general. Paired t-tests revealed that both groups achieved a significant improvement in overall production (experimental group: t (29) = -13.309, $p = .000 < .001$; control group: t (29) = -11.861, $p = .000 < .001$), and a significant difference was also observed in the improvements between the groups (t (58) = 4.993, $p = .000 < .001$). Two-way mixed measures AVOVA with time (pre- and post-test) as the within-subjects factor and groups as the between-subjects factor determined that there were strong interaction effects (F (1,38) = 13.072, p=.000, < .001).

Figure 4

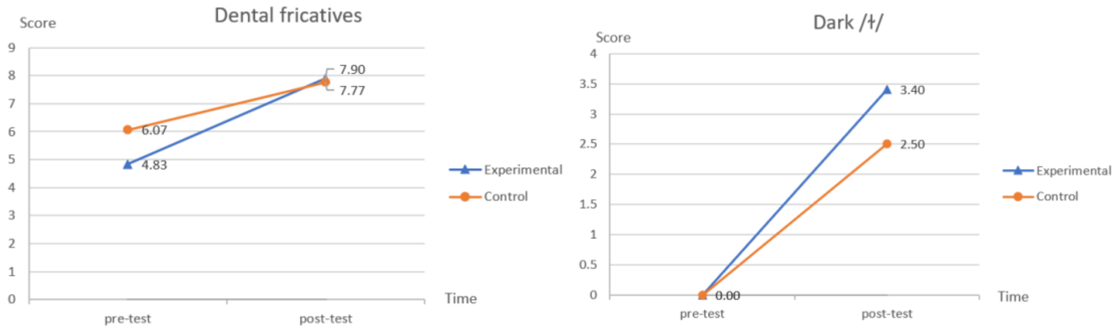
Mean Scores Across the Production Tasks at Pre- and Post-test



Regarding the word-reading task, t-tests showed a significant difference in scores across time for both groups (experimental group: *t* (29) = -12.537, *p* = .000 < .001; experimental group: *t* (29) = -9.265 *p* = .000 < .001). The improvement of the experimental group was significantly higher than the control group in general (*t* (58) = 3.301, *p* = .002 < .01). When we examined each consonant individually, no significant difference between the groups was revealed for the dental fricatives /θ/ and /ð/ from the independent samples t-test (*t* (58) = 1.944, *p* = .057 > .05), since both groups achieved almost perfect scores in the post-test, while the data for dark /ɫ/ demonstrated a significant difference between the groups (*t* (29) = 2.546 *p* = .014 < .05, see Figure 5).

Figure 5

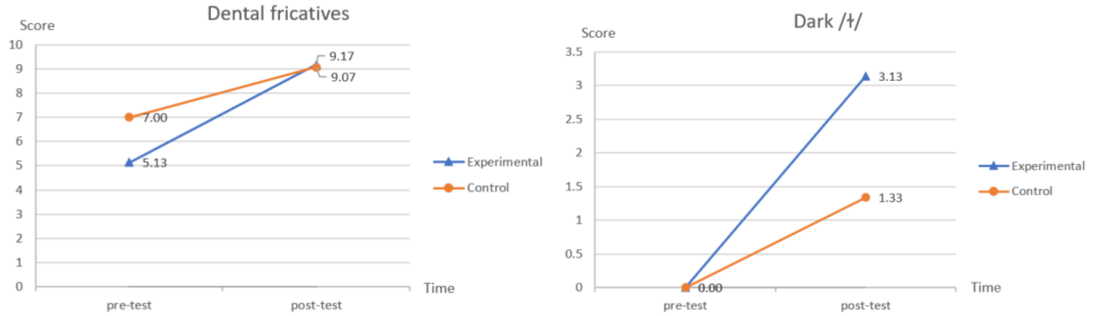
Mean Scores for the Word-reading Task at Pre- and Post-test



In contrast, the analysis of the data from the impromptu translation task at pre- and post-test revealed significant differences in terms of both time and groups (see Figure 6). Paired t-tests based on pre- and post-test results indicated a significant difference between the two tests for both groups (experimental group: *t* (29) = -11.232, *p* = .000 < .001; control group: *t* (29) = -7.999, *p* = .000 < .001). The experimental group improved much more than the control group in their productions of dental fricatives (*t* (58) = 2.702, *p* = .009 < .05) as well as dark /ɫ/ (*t* (58) = 4.584, *p* = .000 < .001). One thing to note is that the difference between the groups in the translation task for dark /ɫ/ (*p* < .001) is more significant than the difference in the word-reading task (*p* < .05). This suggests that the experimental group improved significantly in both tasks. Since impromptu translation can elicit more natural speech than word-reading tasks, the statistics may show that the intervention did help the learners improve their production of the most difficult dark /ɫ/ in natural speech.

Figure 6

Mean Scores for Impromptu Translation Task at Pre- and Post-test



Identification Tests

Paired t-tests showed that both groups demonstrated a significant effect of the time variable in their identification scores (experimental group: *t* (29) = -2.063, *p* = .048 < .05; control group: *t* (29) = -5.979 *p* = .000 < 0.001, see Table 2). The improvement in the control group was more significant than that in the experimental group. The two groups were comparable at pre-test, but the mean score for the control group was slightly higher (1.64 points, or 10.49%) at post-test. The analysis indicated that the interventions were effective in both groups in terms of perception, but the traditional method was more effective.

Table 2

Paired T-test Results for Identification Test

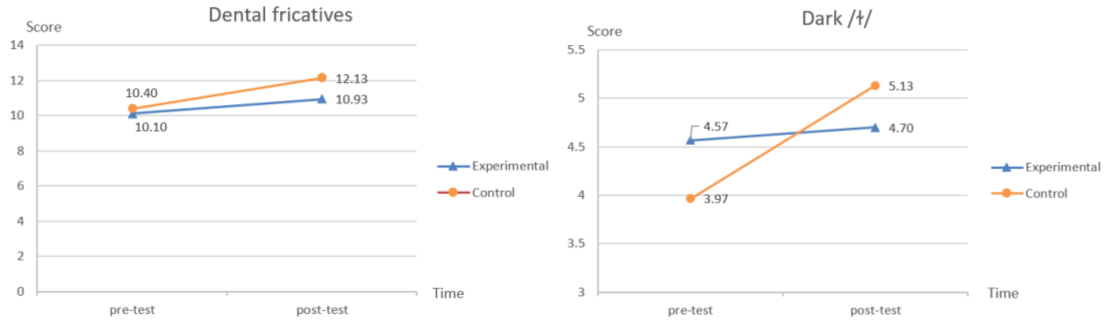
	Pre-test			**Post-test**				
	N	*M*	*SD*	*N*	*M*	*SD*	*t*	*p*
Experimental group	30	14.67	2.68	30	15.63	2.78	-2.063	.048*
Control group	30	14.37	2.91	30	17.27	3.05	-5.979	.000***

Note. * indicates *p* < .05; ***indicates *p* <.001

Considering each sound separately, the left graph in Figure 7 demonstrates that the improvements in both groups were comparable for the two dental fricatives. There was a significant difference between time in both groups (experimental group: *t* (29) = -2.019, *p* = .05; control group: *t* (29) = -3.812, *p* < .001), indicating that both teaching methods were effective for the auditory perception of dental fricatives. Regarding dark /ɫ/ (the right graph in Figure 7), a significant improvement was found for the control group (*t* (29) = -4.051, *p* = .000 < .001), but not for the experimental group (*t* (29) = -4.07, *p* = .687 > 0.05). It therefore seems that the training exerted positive effects for both groups in terms of perception, but the traditional method was more effective, especially for dark /ɫ/.

Figure 7

Mean Score Changes for Dental Fricatives and Dark /ɫ/ Separately in the Identification Tests



Attitude Questionnaire

The first part of the questionnaire aimed to test six factors: instructional approach, learning quality, learning willingness, learning anxiety, learning difficulties, and learning interaction. The Cronbach's alpha coefficient indicated high reliability of the questionnaire (α = .888 for the positive items, α = .786 for the reverse-scoring items). The averages of the six factors in the questionnaire were between 3.59 and 3.98, suggesting an overall positive attitude from all participants towards the application. The results of this part of the questionnaire are presented in Table 3.

Table 3

Mean Scores and Standard Deviations of the 6 Factors in the Questionnaire

Factors	*M*	*SD*
Instructional approach	3.93	.50
Learning quality	3.98	.62
Learning willingness	3.80	.75
Learning anxiety	3.59	.80
Learning difficulties	3.83	.56
Learning interaction	3.91	.69

In response to the final open-ended question, participants reported many suggestions, attitudes, or overall perceptions towards the whole CF process, in addition to the factors discussed above. Of the 30 answers, 25 showed overall positive attitudes towards this AR-enabled online CF model, and looked forward to receiving similar phonetic courses in the future. Four students also provided the explanation: "*The enlarged visual feedback of our lips and tongues with annotations facilitated the recognition of both our own and the teacher's mouth shapes*"; "*In this way, my attention was successfully and quickly attracted to the mouth of the student who was being corrected by the teacher*." Also, one student mentioned that this technology helped him learn more about other students' performances and compare them with his own. Moreover, one student wrote that she loved the use of this funny AR mask which made her less nervous when asked to pronounce the target sound in the class. However, one-third of the respondents reported technical problems, such as insufficient network speed and the installation issue of Snap Camera caused by incompatible computer systems, which required a lot of preparation time. One student thereby expressed the hope that the videoconferencing system might develop its own digital zoom-in function and AR annotation for the camera. "*Thus, it will save us time on software installation and reduce the technical problems*," he said.

Three of 30 participants wrote that they still preferred the offline interaction mode between peers which seemed more active and convenient, and thus they suggested that the AR technologies should be integrated into offline multimedia language classrooms, or that the phonetics course should adopt both online and offline mixed teaching modes.

Discussion

This study set out to address three research questions: 1) the effects of multimodal CF through AR-enabled videoconferencing on L2 production of the target segments; 2) the impact of the new CF model on L2 perception of the target segments; and 3) learners' attitudes towards the new CF environment compared to traditional offline CF.

With regards to RQ1, the statistical analysis revealed that although both groups improved significantly in the post-test word-reading production task ($p = .000 < .001$) and achieved almost perfect scores for the dental fricatives /θ/ and /ð/, the experimental group improved significantly more than the control group in the production of dark /ɫ/ ($p = .014 < .05$), which was shown by our pre-test scores to be more difficult for Mandarin speakers to pronounce correctly than the dental fricatives (45.42% accuracy for /θ/ and /ð/ but zero for dark /ɫ/). The articulatory analysis shows that the dental /θ/ and /ð/ are more visually salient than the alveolar dark /ɫ/. In traditional offline teaching settings, the articulatory gesture of dental fricatives—putting the tongue tip between the upper and lower teeth—can be easily detected by learners as long as the teacher demonstrates the frontal view of his/her mouth or those of the model speakers on the screen. However, the alveolar dark /ɫ/ requires the speakers to lift the tongue tip to the alveolar ridge in the mouth. Therefore, the visual recognition of the position of the tongue tip in the mouth and noticing the lip shape is different from the rounded lip for /ʊ/ are essential preconditions of its correction. The results of the production tests indicate that the digital zoom-in lens with AR annotation plugged in the videoconferencing system is very helpful in improving the controlled production of the most difficult and less visually salient dark /ɫ/ as they provide the enlarged and annotated real-time images of the articulators in the process of CF.

Moreover, in the spontaneous translation task the experimental group also improved significantly for all the target segments, whereas the control group did not. Interestingly, this result partially echoes the view of Rau et al. (2009) which was based on a classroom study showing that Chinese learners of English mispronounced /θ/ more frequently in a picture description task than in word and sentence reading tasks. According to DeKeyser (2017) and Saito and Plonsky (2019, p. 665), in the field of pronunciation teaching research "the effect of instruction in the initial stages of L2 speech learning (i.e., noticing and consolidation of declarative knowledge)" can be revealed through the examination of performance in controlled speech tasks, while "examining spontaneous speech performance is assumed to reveal the role of instruction in the mid and later stages of L2 speech learning (i.e., proceduralization and automatization)" (Saito & Plonsky, 2019, p.13). Given that the new multimodal CF model in this study supported the processes of receiving (audiovisual input), interacting, and speech production (verbal output)—all three of the crucial SLA functions—the significant difference between the performances of the experimental and control groups in the spontaneous task may demonstrate that the online multimedia CF model plays a greater role in facilitating the proceduralization of articulatory knowledge compared to traditional offline CF. In addition, since the AR annotation and filters, which are vital components of our new CF model, are functioning as gesture-based visual reinforcement tools, this finding was also in line with the results of previous empirical studies suggesting that exposure to articulatory gestures may be effective in improving learners' acquisition of L2 sounds that they initially have difficulty with (Hazan et al., 2005; Li & Somlak, 2017).

In relation to RQ2, the results of the identification tests show that although both groups improved significantly in the identification of all the target segments, the control group improved much more than the experimental group in the auditory identification tasks for both dental fricatives and dark /ɫ/. This finding demonstrates that both the CF environments (i.e., the new online multimodal CF and traditional offline CF) were able to improve learners' auditory identification of L2 segments. It is worth noting that the CF environments for both the control and experimental groups involved audiovisual multimedia and the new online model was equipped with visual articulatory enhancement tools, namely AR and videoconferencing.

The natural teaching scenarios were different from the controlled laboratory conditions (audiovisual condition vs. audio-only condition) in previous studies (Navarra & Soto-Faraco, 2007; Okuno & Hardison, 2016) in terms of the facilitating effects of visual feedback on L2 perception and production. Visual articulatory reinforcement may be a helpful but not necessary condition for the improvement of auditory perception. The results of identification tests indicate that the existing offline multimedia environment in favor of audio support in both input and interaction is more advantageous in the improvement of auditory identification of L2 segments contrasts. Meanwhile, gesture-based visual information seemed to create a split-attention effect in the training of learners' ears, since it directed more of their attention to articulatory gestures. This finding is consistent with the results of previous empirical research on the issue of L2 output within a multimedia environment (Chun & Plass, 1996; Chun, 2001; Jones, 2004), which suggest that "in multimedia learning environments, teaching modes and test modes should be compatible" (Plass & Jones, 2005, p. 467).

An unexpected result was shown in the pre-tests for the most difficult segment, dark /ɫ/, which none of the participants pronounced correctly, but for which the accuracy rate in the identification test was 53.33%. This might imply that learners can perceive L2 sounds before production. Combined with all the results from the production tests, this finding may provide indirect evidence for the hypothesis that there is no strong correlation between how accurately late learners produce and perceive L2 phonetic segments (Flege, 1999). It is also consistent with Strange's (1995) conclusion that perceptual difficulties may still persist even after segmental production has been mastered. Our analysis of the data collected in the production and identification tests revealed that the use of AR-enabled videoconferencing in the online CF class was beneficial to learners' correction of their segmental production even before they achieved accuracy in the auditory perception of those sounds. However, whether this production competency may be transferred to a significant improvement of perception in the long run, especially after full automatization of the relevant articulatory knowledge, is still a missing piece of the puzzle.

Finally, with regard to RQ3, the quantitative data collected from the questionnaire reflected a generally positive attitude from the participants towards the application of the new CF model in place of traditional offline CF. The qualitative data collected from the answers to the final open-ended question might offer a reasonable explanation for the connection between the application of the new CF model and the improvement in the learners' segmental production. The majority of them felt satisfied with the learning quality of the new model, because the AR-enabled videoconferencing guided their attention to the articulators of the student being corrected, so that they were less easily distracted during the one-on-one pronunciation CF. The real-time AR annotation in this study could be seen as visual cues added to the mirror of the leaners' articulatory gestures. As van Gog (2014) claimed, this high-cueing condition could have a positive effect on cognitive load since it guided learners' attention to the essential material. Furthermore, the AR filter used in this study helped to remove other facial features which were irrelevant to pronunciation learning, reducing the extraneous cognitive load that was ineffective for learning or may even hamper learning (Sweller et al., 2011). Thus, learning efficiency was reported to be greatly enhanced. Since the audiovisual model selected from the IPA Phonetics video corpus and the learner's articulation to be corrected could be displayed on one screen via videoconferencing, the participants found it easier to identify mismatches between the model and their own production. These results not only corroborated with those from the production tests, but also provided evidence for the utility of the coherence, signaling, spatial contiguity, and temporal contiguity principles on the basis of which our AR and videoconferencing tools were founded on and organized to minimize extraneous overload in multimedia teaching environments (Mayer & Fiorella, 2014). Given these advantages, most participants expressed their willingness to receive this type of CF in the future.

However, the qualitative data from the final open-ended question also revealed issues with technical and social affordances. Although they favored the new CF model, three participants gave advice on how to avoid these possible problems and improve the online class condition. First, they suggested that the AR technology should be integrated in a group videoconferencing function to encourage more group interactions, such as peer CF. Second, some students need to be trained further on how to use technologies, especially skills related to peer interaction. Finally, the AR technology should be further developed and applied in offline language classrooms. This integration may not only enhance the efficiency of class CF but also guarantee active peer interaction.

Conclusion and Pedagogical Implications

Recent years have seen significant developments in studies of intelligent feedback in pronunciation training due to technology such as automatic speech recognition (ASR) (Hardison, 2004; Olson, 2014). However, most of these studies have been restricted to laboratory settings and have focused on the effectiveness of human-computer interaction in pronunciation training. The present study aimed to investigate how the integration of AR technology with a videoconferencing platform could facilitate teacher-student CF interaction in L2 pronunciation classrooms.

The quasi-experimental data showed that the online multimodal CF environment equipped with AR annotation and filters played a significant role in improving the participants' production of the target segments, especially the most difficult dark /ɫ/ for Mandarin speakers. However, it offered no advantage in improving their auditory perception compared to the offline CF class using traditional multimedia aids, since the multimodal CF method that we designed was intended to provide visual cues to reduce multimedia overload and focus on the essential information. In addition, the quantitative and qualitative analyses of the questionnaire both revealed the general positive attitudes of all the participants. In summary, the employment of the new technology was not only effective in terms of segmental production, but also helped to draw learners' attention to the articulatory gestures of the student being corrected, and motivated them to participate fully in class interactions (e.g., teacher's CF and peer CF).

The findings of the present study are significant to SLA using multimedia in several ways. First, a particular strength of the study lies in our deliberate choice of design principles based on multimedia learning rationale. The employment of synchronous videoconferencing in our L2 pronunciation CF class was not driven by the COVID-19 pandemic situation, as a passive choice for emergency reasons. As Fuchs (2022) notes, there is a distinguishing difference between emergency remote teaching and online teaching, since educators in emergency teaching situations may not have had the time and experience to adjust course syllabi to guarantee or enhance teaching efficacy. This study designed, organized, and integrated multimedia materials in order to solve the cognitive overload problems inherent in L2 pronunciation CF class with traditional multimedia aids. Both the teacher and students should be well-trained and prepared for the application of advanced technology.

Second, the results indicate that different teaching modes should be applied for different modes of tasks, such as production and auditory perception. For example, visual articulatory reinforcement tools should only be used to improve segmental production in the CF link, while we can turn to traditional offline language classrooms or adopt other auditory reinforcement tools to solve perceptual problems.

Third, AR technology providing virtual elements superimposed upon or composited with the real world is usually seen as an addition to traditional multimodal teaching environments. However, grounded in the cognitive theory of multimedia learning, this study treated AR technology as a method to reduce the cognitive overload caused by too much multisensory and multimodal information. AR annotation and filters were used to draw learners' attention to the essential L2 articulatory information and reduce extraneous cognitive load, enhancing CF efficiency. In this sense, this type of subtraction may be the best addition to L2 learning.

Furthermore, this study created a pronunciation CF model through AR-enabled videoconferencing by aligning AR technology affordances with design principles for multimedia learning. This alignment offers new theoretical and pedagogical implications, since there is a dearth of empirical studies on how to use AR based on a theoretical grounding to deliberately support the SLA of specific skills (Parmaxi & Demetriou, 2020). Last but not least, the implementation of AR-enabled videoconferencing synchronized and aligned learners' real-time non-target production with videotaped production models, not only enhancing the teacher's and learners' cognitive processing in both time and space, but also restructuring the traditional classroom ecology. In this sense, this AR application is not only beneficial for the development of CAPT, but also provides a new inspiration for the design and construction of metaverse classrooms.

One limitation of the study is the lack of delayed post-tests for either group in the testing procedure, in order to avoid potential effects of subsequent teaching and CF for other segments in the pronunciation course. Another limitation is the identification test which only investigated auditory perception. Since the CF process was conducted in a dynamic multimedia context, in future studies, the corresponding identification tests should also be multimodal, involving audiovisual presentation of stimuli, which may reach different perceptual results.

Acknowledgements

We would like to show our gratitude to the participants and judges. Also, we would 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].

References

Agustuna, N. E., Herlina, R., & Faridah, D. (2019). Corrective feedback on pronunciation errors: teacher's perception and EFL high school students' self-reflection. *Journal of English Education and Teaching, 3*(3), 311–327. https://doi.org/10.33369/jeet.3.3.311-327

Azuma, R. T. (1997). A survey of augmented reality. *Teleoperators and Virtual Environments, 6*(4), 355–385. https://doi.org/10.1162/pres.1997.6.4.355

Badin, P., Tarabalka, Y., Elisei, F., & Bailly, G. (2010). Can you 'read' tongue movements? Evaluation of the contribution of tongue display to speech understanding. *Speech Communication, 52*(6), 493–503. https://doi.org/10.1016/j.specom.2010.03.002

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. (1994). The emergence of native-language phonological influences in infants: A perceptual assimilation model. In J. C. Goodman & H. C. Nusbaum (Eds.), *The development of speech perception: The transition from speech sounds to spoken words* (pp. 167–224). MIT Press.

Best, C. T. (1995a). A direct realist view of cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience* (pp. 171–204). York Press.

Best, C. T. (1995b). Learning to perceive the sound pattern of English. In C. Rovee-Collier & L. Lipsitt (Eds.), *Advances in infancy research* (pp. 217–304). Ablex.

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In M. Munro & B. Ocke-Schwen (Eds.), *Language Experience in Second Language Speech Learning: In Honor of James Emil Flege* (pp. 13–34). John Benjamins.

Carmigniani, J., & Furht, B. (2011). Augmented reality: An overview. In B. Furht (Ed.), *Handbook of augmented reality* (pp. 3–46). Springer.

Chun, D. M. (2001). L2 reading on the web: Strategies for accessing information in hypermedia. *Computer Assisted Language Learning, 14*(5), 367–403. https://doi.org/10.1076/call.14.5.367.5775

Chun, D. M., Jiang, Y., Meyr, J. M., & Yang, R. (2015). Acquisition of L2 Mandarin Chinese tones with learner-created tone visualizations. *Journal of Second Language Pronunciation, 1*(1), 86–114. https://doi.org/10.1075/jslp.1.1.04chu

Chun, D. M., & Plass, J. L. (1996). Effects of multimedia annotations on vocabulary acquisition. *The Modern Language Journal, 80*(2), 183–198. https://doi.org/10.1111/j.1540-4781.1996.tb01159.x

Correia, A., Liu, C., & Xu, F. (2020). Evaluating videoconferencing systems for the quality of the educational experience. *Distance Education, 41*(4), 429–452. https://doi.org/10.1080/01587919.2020.1821607

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.

de Koning, B. B., Tabbers, H. K., Rikers, R. M. J. P., & Paas, F. (2009). Towards a framework for attention cueing in instructional animations: Guidelines for research and design. *Educational Psychology Review, 21*, 113–140. https://doi.org/10.1007/s10648-009-9098-7

Deterding, D. (2006). The pronunciation of English by speakers from China. *English World-Wide, 27*(2), 175–198. https://doi.org/10.1075/eww.27.2.04det

Deterding, D. (2007). Singapore English (Dialects of English). Edinburgh: Edinburgh University Press.

Ding, S., Liberatore, C., Sonsaat, S., Lučić, I., Silpachai, A., Zhao, G., Chukharev-Hudilainen, E., Levis, J., & Gutierrez-Osuna, R. (2019). Golden speaker builder – An interactive tool for pronunciation training. *Speech Communication, 115*, 51–66. https://doi.org/10.1016/j.specom.2019.10.005

Dlaska, A., & Krekeler, C. (2013). The short-term effects of individual corrective feedback on L2 pronunciation. *System, 41*(1), 25–37. https://doi.org/10.1016/j.system.2013.01.005

Engwall, O. (2008). Can audio-visual instructions help learners improve their articulation? - An ultrasound study of short-term changes. In *Proc. Interspeech* 2008, 2631–2634, https://doi.org/10.21437/Interspeech.2008-652

Engwall, O. (2012). Analysis of and feedback on phonetic features in pronunciation training with a virtual teacher. *Computer Assisted Language Learning, 25*(1), 37–64. https://doi.org/10.1080/09588221.2011.582845

Flege, J. (1995). Second language speech learning: Theory, findings and problems. In W. Strange (Ed.), *Speech perception and linguistic experience* (pp. 233–272). York Press.

Flege, J. E. (1999). The relation between L2 production and perception. *XIV International Congress of Phonetic Sciences, 2*, 1273–1276.

Fouz-González, J. (2015). Trends and directions in computer assisted pronunciation training. In J. Mompean & J. Fouz-González (Eds.), *Investigating English Pronunciation: Trends and directions* (pp. 314–342). Palgrave Macmillan.

Fuchs, K. (2022). The difference between emergency remote teaching and e-learning. *Frontiers in Education, 7*, 1–3. https://doi.org/10.3389/feduc.2022.921332

Hardison, D. M. (2003). Acquisition of second-language speech: Effects of visual cues, context, and talker variability. *Applied Psycholinguistics, 24*(4), 495–522. https://doi.org/10.1017/S0142716403000250

Hardison, D. M. (2004). Generalization of computer-assisted prosody training: Quantitative and qualitative findings. *Language Learning & Technology, 8*(1), 34–52. http://dx.doi.org/10125/25228

Havranek, G. (2002). When is corrective feedback most likely to succeed*? International Journal of Educational Research*, *37*(3-4), 255–270. https://doi.org/10.1016/S0883-0355(03)00004-1

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., & Nguyen, P. (2021). Technology-mediated corrective feedback. In H. Nassaji & E. Kartchava (Eds.), *The Cambridge Handbook of Corrective Feedback in Second Language Learning and Teaching* (pp. 226–250). Cambridge University Press.

Jones, L. C. (2004). Testing L2 vocabulary recognition and recall using pictorial and written test items. *Language Learning & Technology, 8*(3),122–143. http://dx.doi.org/10125/43998

Kartchava, E., & Ammar, A. (2013). Learners' beliefs as mediators of what is noticed and learned in the language classroom. *TESOL Quarterly, 48*(1), 86–109. https://doi.org/10.1002/tesq.101

Khoroshilova. S. (2016). Anxiety in a foreign language pronunciation class in a university setting. In J. Przedlacka, Maidment & Ashby (Eds.), *Proceedings of PTLC2013, Papers from the Phonetics Teaching and Learning Conference* (pp. 541–548). London: PTLC; 19-22.

Kim, J., & Han, Z. (2007). Recasts in communicative EFL classes: Do teacher intent and learner interpretation overlap? In A. Mackey (Ed.), *Conversational Interaction in Second Language Acquisition: A Collection of Empirical Studies* (pp. 269–297). Oxford University Press.

Klimova, B., & Pikhart, M. (2022). Application of corrective feedback using emerging technologies among L2 university students. *Cogent Education, 9*(1), 1–14. https://doi.org/10.1080/2331186X.2022.2132681

Küçük, S., Yılmaz, R. M., Baydaş, Ö., & Göktaş, Y. (2014). Augmented reality applications attitude scale in secondary schools: Validity and reliability study. *Education & Science, 39*(176), 383–392. https://doi.org/10.15390/EB.2014.3590

Lee, E. J. (2016). Reducing international graduate students' language anxiety through oral pronunciation corrections. *System, 56*, 78–95. https://doi.org/10.1016/j.system.2015.11.006

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, 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

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of speech code. *Psychological Review, 74*(6), 431–461. https://psycnet.apa.org/doi/10.1037/h0020279

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21*(1), 1–36. https://doi.org/10.1016/0010-0277(85)90021-6

Liberman, A. M., & Whalen, D. H. (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

Liu, Y., & Tseng, W. (2019). Optimal implementation setting for computerized visualization cues in assisting L2 intonation production. *System, 87*, 102–145. https://doi.org/10.1016/j.system.2019.102145

Lyster, R., Saito, K., & Sato, M. (2013). Oral corrective feedback in second language classrooms. *Language Teaching, 46*(1), 1–40. https://doi.org/10.1017/S0261444812000365

Mackey, A. (2006). Feedback, noticing and instructed second language learning. *Applied Linguistics, 27*(3), 405–430. https://doi.org/10.1093/applin/ami051

Martin, I. A. (2020). Pronunciation development and instruction in distance language learning. *Language Learning & Technology, 24*(1), 86–106. https://doi.org/10125/44711

Massaro, D. W. (1998). *Perceiving talking faces.* The MIT Press.

Mayer, R. (Ed.). (2014). *The Cambridge Handbook of Multimedia Learning* (2nd ed.). Cambridge University Press.

Mayer, R. E., & Fiorella, L. (2014). Principles for reducing extraneous processing in multimedia learning: Coherence, signaling, redundancy, spatial contiguity, and temporal contiguity principles. In R. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning (2nd ed.)* (pp. 279–315). Cambridge University Press.

Mayer, R., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist, 38*(1), 43–52. https://doi.org/10.1207/S15326985EP3801_6

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

Okuno, T., & Hardison, D. M. (2016). Perception-production link in L2 Japanese vowel duration: Training with technology. *Language Learning & Technology, 20*(2), 61–80. http://dx.doi.org/10125/44461

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

Parmaxi, A., & Demetriou, A. A. (2020). Augmented reality in language learning: A state-of-the-art review of 2014–2019. *Journal of Computer Assisted Learning, 36*(4), 861–875. https://doi.org/10.1111/jcal.12486

Patten, I., & Edmonds, L. A. (2015). 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

Pawlak, M., Zawodniak, J., & Kruk, M. (2021). Individual trajectories of boredom in learning English as a foreign language at the university level: Insights from three students' self-reported experience. *Innovation in Language Learning and Teaching, 15*(3), 263–278. https://doi.org/10.1080/17501229.2020.1767108

Plass, J. L., & Jones, L. C. (2005). Multimedia learning in Second Language Acquisition. In R. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning (1st ed.)* (pp. 467–488). Cambridge University Press.

Rau, D., Chang, A., & Tarone, E. (2009). Think or sink: Chinese learners' acquisition of the voiceless interdental fricative. *Language Learning, 59*(3), 581–621. https://doi.org/10.1111/j.1467-9922.2009.00518.x

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. (2015). Variables affecting the effects of recasts on L2 pronunciation development. *Language Teaching Research, 19*(3), 276–300. https://doi.org/10.1177/1362168814541753

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

Schulz, R. A. (2001). Cultural differences in student and teacher perceptions concerning the role of grammar instruction and corrective feedback: USA—Colombia. *The Modern Language Journal, 85*(2), 244–258. https://doi.org/10.1111/0026-7902.00107

Sommer, R. (1967). Classroom ecology. *The Journal of Applied Behavioral Science, 3*(4), 489–503. https://doi.org/10.1177/002188636700300404

Strange, W. (1995). Phonetics of second-language acquisition: Past, present, future. In K. Elenius & P. Branderud (Eds.), *Proceedings of the XIIIth International Congress of Phonetic Sciences* (pp. 76–84). Stockholm University.

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory.* Springer.

Talidong, K. J. B. (2020). Implementation of emergency remote teaching (ERT) among Philippine teachers in Xi'an, China. *Asian Journal of Distance Education, 15*(1), 196–201. https://eric.ed.gov/?id=EJ1290051

Thomson, R. I. (2011). Computer assisted pronunciation training: Targeting second language vowel perception improves pronunciation. *CALICO Journal, 28*(3), 744–765. https://eric.ed.gov/?id=EJ956341

Tsai, P. H. (2019). Beyond self-directed computer-assisted pronunciation learning: A qualitative investigation of a collaborative approach. *Computer Assisted Language Learning, 32*(7), 713–744. https://doi.org/10.1080/09588221.2019.1614069

van Gog, T. (2014). The signaling (or cueing) principle in multimedia learning. In R. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning (2nd ed.)* (pp. 263–278). Cambridge University Press.

van Wassenhove, V., Grant, K. W., & Poepel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences of the United States of America, 102*(4), 1181–1186. https://doi.org/10.1073/pnas.0408949102

Wang, Y. (2004). Supporting synchronous distance language learning with desktop videoconferencing. *Language Learning & Technology, 8*(3), 90–121. http://dx.doi.org/10125/43997

Wiesemes, R., & Wang, R. (2010). Video conferencing for opening classroom doors in initial teacher education: Sociocultural processes of mimicking and improvisation. *International Journal of Media, Technology and Lifelong Learning, 6*(1), 28–42. https://doi.org/10.7577/seminar.2456

Wu, H.-K., Lee, S. W.-Y., Chang, H.-Y., & Liang, J.-C. (2013). Current status, opportunities and challenges of augmented reality in education. *Computers & Education, 62*, 41–49. https://doi.org/10.1016/j.compedu.2012.10.024

Yeh, H.-C. & Tseng, S.-S. (2020). Enhancing multimodal literacy using augmented reality. *Language Learning & Technology, 24*(1), 27–37. https://doi.org/10125/44706

Zhu, J., Zhang X., & Li, J. (2022). Using AR filters in L2 pronunciation training: Practice, perfection, and willingness to share. *Computer Assisted Language Learning*, published online, https://doi.org/10.1080/09588221.2022.2080716

Appendix A. Production Tests

1. Tokens for the word-reading task:

/θ/	**th**ief	**th**ank	weal**th**	**th**eme
/ð/	clo**th**	wi**th**	brea**th**e	**th**ere
dark l	litt**le**	arriv**al**	de**al**	e**l**se

2. Translation task: Please interpret each of the following Chinese sentences into English in 10 seconds.

1) 他们把苹果放在了那个书架上。

Answer: They put the apple on that shelf.

(target words: **th**ey, app**le**, she**l**f)

2) 真相是，这个小偷翻墙后逃走了。

Answer: The truth is that the thief climbed over the wall and fled.

(target words: tru**th**, **th**ief, wa**ll)**

3) 碗里装着的牛奶不多了。

Answer: Little milk is left in the bowl.

(target words: bow**l**, mi**l**k)

4) 我把嘴巴张开，看到我第四颗牙齿坏掉了。

Answer: I open my mouth to find the fourth teeth is broken.

(target words: mou**th**, four**th**, tee**th)**

5) 妈妈用冷水洗了个澡。

Answer: My mum takes a shower with cold water.

(target words: co**l**d)

6) 今天天气很好，我和哥哥一起出去逛街

Answer: The weather is nice today, and I go out shopping with my brother.

(target words: wea**th**er, wi**th**, bro**th**er)

Appendix B. Identification Tests

1. Listen to the audio carefully, write down "T" if the word is correctly pronounced, or "F" if the word is mispronounced.

/θ/	**th**reat	**th**ink	**th**row	ba**th**	tee**th**
/ð/	smoo**th**	**th**en	brea**th**e	clo**the**	**th**ough
dark /ɫ/	a**ll**	cool	so**l**d	fu**ll**	muff**le**

2. Choose the right phonetic symbol according to what you've heard:

		A	**B**	**C**
/θ/	1)	/'srufri/	/'θrufri/	
	2)	/bres/	/breθ/	
	3)	/'banles/	/'banleθ/	
/ð/	4)	/'ðepna/	/'zepna/	/'depna/
	5)	/'ðailish/	/'zailish/	/'dailish/
	6)	/gro'nʌð/	/gro'nʌz/	/gro'nʌd/
dark /ɫ/	7)	/'kreshful/	/'kreshfɔ:/	
	8)	/lul/	/lu:/	/luə/
	9)	/'prutl/	/'prutɔ:/	/'prutəu/

Appendix C. Questionnaire

Themes	**Questions**	**English Translation**
Instructional approach	1. 腾讯会议中的 AR 插件对我的学习来说很有帮助。	1. The AR plug-in in Tencent Meeting was very helpful to learning pronunciation.
	2.相比之前，我很喜欢这种教学模式。	2. I prefer the lessons instructed with this approach rather than the traditional one.
	3.这种教学模式下我很难跟上课堂节奏。	3. It was difficult to follow the CF model because I found it confusing.

Learning quality	1. 这种包含 AR 应用的教学模式能帮助我更好的集中注意力。	1. I could concentrate better on the lesson when the AR applications were used.
	2.这种学习模式大大提高了我的学习效率。	2. This CF model has largely enhanced my learning efficacy.
	3.对于纠音而言，这种教学模式没有带来任何好处。	3. The model has brought no additional benefits to error correction regarding pronunciation.
Learning willingness	1.如果我还会上语音课，我希望它也能采取这种教学模式。	1. If I take another phonetic course, I hope the AR applications will be adopted in those lessons as well.
	2.我希望语音课能更多的采用线上教学而非线下。	2. I hope that phonetic courses should be instructed online instead of offline.
	3.我不希望以后的课程中出现这种教学模式。	3. I don’t want to follow this model anymore in future courses.
Learning anxiety	1. 相比以前的课程，我在这堂课上被老师纠正时没有那么紧张了。	1. I don’t feel so stressed when being corrected in class by the teacher as I used to be.
	2. 我很担心我发音错误时别人对我的看法。	2. I am very worried about what others think of me when mispronouncing a word.
	3. 老师让我打开摄像头时我感到很尴尬。	3. I felt embarrassed when the teacher asked me to switch on the camera in the class.
Learning difficulties	1.我觉得这种模式下，擦音（/θ/和/ð/）和舌侧音（dark l）的纠正变得更简单了。	1. The correction of /θ/, /ð/ and dark /ɫ/ became easier in this new CF model.
	2.这种教学模式降低了纠音的难度。	2. This teaching approach has lowered the difficulty of correcting pronunciation errors.
	3.我在上课期间遇到了很多问题，例如设备问题、线上教学的沟通问题等。	3. I encountered many problems during class (e.g. technical problems, communication problems, etc.).
Learning interaction	1.观察老师纠正其他同学的发音时，我觉得我的发音也有提高。	1. Through observing my peers being corrected by the teacher, I found that my pronunciation improved too.

	2. 课堂上的互动环节提高了课堂效率。	2. The interactive activities in classes have improved classroom efficacy.
	3.互动环节对我来说是个负担。	3. The interactive activity was a burden for me.

About the Authors

Yiran Wen is an undergraduate student pursuing her bachelor degree at the School of Foreign Studies, Shanghai University of Finance and Economics. Her research interests include phonetics, CAPT, and multimedia learning.

E-mail: alison_wen@foxmail.com

Jian Li (corresponding author) is an associate professor at the 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.

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

Hongkang Xu is an undergraduate student in Computer Science and Cognitive Psychology at Northeastern University. His research interests include artificial intelligence, machine learning, computer vision, and social behavior.

E-mail: xu.hong@northeastern.edu

Hanwen Hu is an undergraduate student in Business English at Shanghai University of Finance and Economics. His research interests include applied linguistic and technology-assisted language learning.

E-mail: oliverhusufe@gmail.com