Article

LANGUAGE
TEACHING
RESEARCH

Language Teaching Research
1–15
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/13621688231219773
journals.sagepub.com/home/ltr
Sage

Enhancing L2 sound learning through the integration of audio-visual information: Phonetic training in the classroom

Ying Li
City University of Macau, China

Abstract

The present study aimed to investigate the effectiveness of an audio-visual high variability phonetic training (AV-HVPT) approach in improving the pronunciation of English interdental sounds /θ, ð/ and vowel /ɪ/ among a group of Chinese graduates. A total of 70 participants were randomly assigned to two groups: one received AV-HVPT instruction (Class 1) and the other one received traditional teacher-led instruction (Class 2). The overall findings showed that Class 1 significantly improved /θ, ð/ pronunciation but Class 2 did not. No improvement in /ɪ/ pronunciation was observed in either group. This may suggest that the effectiveness of the AV-HVPT approach in the classroom might be limited in the learning of sounds with salient articulatory gestures.

Keywords

adult learners, audio-visual training, English sounds, high variability phonetic training (HVPT), L2 production

I Introduction

Exposure to the second language (L2) in the target community is suggested to be a preferred approach to learn the pronunciation of the L2 (Krashen, 1981), although factors other than learning environment may also intervene in L2 sound learning (Trofimovich et al., 2015). However, in many foreign language learning contexts, most learners are mainly exposed to the L2 through classrooms and mainly rely on teachers' instructions

Corresponding author:
Ying Li, City University of Macau, Macau, China.
Email: liying_22@163.com

and interventions (Barriuso & Hayes-Harb, 2018; Kachru, 1990). Although scholars have made efforts to modify and improve the instructional approach, problems in L2 speakers' intelligibility still remain (Levis, 2020). It is common for learners (particularly adults) to 'incorrectly'[1] pronounce certain L2 sounds despite having learned the L2 at schools for years (Krashen, 1985). Studies on phonetic training conducted in laboratories have revealed that audio-perceptual training following a high-variability approach received certain degrees of success in ameliorating L2 sound pronunciation (Iverson et al., 2005; Lively et al., 1993). More recently, it was revealed that audio-visual training that follows a high-variability phonetic training (AV-HVPT) approach could also be effective in facilitating L2 sound learning (e.g. Hazan et al., 2005; Ortega-Llebaria et al., 2001). However, few studies have applied this approach in real classrooms (Li & Somlak, 2019). Thus, it is of particular interest to study whether an AV-HVPT approach can be successfully adopted in classroom settings. To fill in the gap, the present study applied an AV-HVPT approach in an L2 English course among a group of first language (L1) Chinese graduates for the learning of English sounds.

II Literature review

Visual articulatory information, or visual codes / visual cues of speech sounds, refers to the vocal gestures used to produce sounds (Li, 2016). Speech sounds are predicted to be more accurately perceived when listened to with matched visual articulatory information (Chen, 2001; Hazan et al., 2005; Hirata & Kelly, 2010). Some theories, such as the motor theory of speech perception (Liberman & Mattingly, 1985; Liberman & Whalen, 2000; Liberman et al., 1967) and the direct realist theory (Fowler, 1994a, 1994b, 1996), state that the object of speech perception is to recover the articulatory gestures of speech sounds. Furthermore, seeing a speaker's articulatory information correlated with its auditory speech signal can significantly improve speech comprehension when the auditory signal is compromised (Ross et al., 2007; Sumby & Pollack, 1954). In line with this, phonetic training using the AV approach that provides articulatory information of target sounds has successfully led to improvements in learners' pronunciation and/or perception (e.g. Hazan et al., 2005; Li & Somlak, 2019). For instance, Bernstein et al. (2013) found that in the perception of paired nonsense words and nonsense pictures, the audio-visual (AV) training group displayed significantly higher accuracy rates than the audio-only (AO)-training group and the nontrained group. The AV-training group also significantly outperformed the AO-training group in the identification of consonants embedded in untrained nonsense words.

It could be true that L2 learners may struggle to correlate an auditory speech signal with its corresponding articulatory gestures. For example, it is common for speakers to incorrectly pronounce certain L2 sounds, such as Japanese speakers' mispronunciation of English /ɹ/–/l/ (Bada, 2001; Best & Strange, 1992; Sennema et al., 2003) and French speakers' difficulties in producing English /θ/ (Picard, 2002), which were attributed to the fact that articulatory settings are not activated in the speakers' L1 (Best & Strange, 1992).

However, evidence from theories and models on L2 sounds learning revealed that non-native sounds are learnable in adulthood. For example, The Speech Learning Model (SLM) (Flege, 1995a, 1995b) and the Perceptual Assimilation Model (PAM)

(Best & Tyler, 2007) both suggest that, given sufficient L2 input, adult learners can eventually learn L2 sounds not present in the L1. Particularly, SLM states that language learners' capacity for speech learning remains intact throughout life and that L2 learning benefits from accurate and abundant input; learners may struggle to distinguish certain L2 sounds at the beginning, but new categories of sounds can eventually be established with an increase of L2 experience (Flege, 1995a, 1995b, 2003). PAM-L2, on the other hand, emphasizes the key role of articulatory gestures in L2 sounds learning, and predicts that the degree of learners' success in learning an L2 sound depends on whether and/or to what extent could he/she correctly perceive the sound's articulatory gestures (Best & Tyler, 2007). This is in accordance with Motor Theory and Direct Realist Theory, both of which also indicate that the objective of speech perception is to recover sounds' articulatory gestures (Fowler, 1996; Liberman et al., 1967). The importance of articulatory information in speech perception, therefore, may rationalize the success of adopting AV approach in phonetic trainings.

Moreover, high variability phonetic training (HVPT) has been found to further enhance L2 sound learning (Bradlow et al., 1997; Iverson et al., 2005; Logan et al., 1991, 1993; Pisoni & Lively, 1995). It aims to direct learners' attention towards relevant phonetic cues by providing stimuli of a wide range of naturally produced exemplars (Iverson et al., 2005; Logan et al., 1991, 1993). For instance, Carlet (2017) investigated the effects of HVPT training approaches under two conditions (categorical discrimination and identification) on Spanish–Catalan speakers' perception and production of English vowels (/i ɪ æ ʌ ɜː/) and initial and final stops. The overall findings revealed that both conditions led to the trainees' significant improvement in their identification of trained sounds. Given their positive effects, some scholars combined AV and HVPT approaches in phonetic trainings in laboratory settings (e.g. Aliaga-García & Mora, 2009; Hazan et al., 2005). For example, Li (2016) conducted a 9-session AV-HVPT training programme among 29 Mandarin adults in the learning of English /θ/–/s/ and /ð/–/z/. The participants were exposed to multiple native English speakers' audio-visual recordings that embedded the target contrasts in various phonetic environments. Their perception and production performance were tested before, during and after the training programme and compared against performance of another group which did not receive AV-HVPT training. According to the results, the trained participants' perception and production accuracies in the post-test were significantly higher than that in the pre-test and were significantly higher than those of non-trained participants in the middle- and post-test (Li, 2016). Similarly, Wang et al. (2008) explored the training effects of audio, visual and AV modalities on Mandarin speakers' perception of English interdental fricatives, which are non-existent in Mandarin. Forty-four Mandarin-speaking adults were randomly assigned to experience audio training, visual training, AV training or not being trained at all (the control group). Percent correct identification of the target sounds at pre-test and post-test revealed significant beneficial effects of AV training on L2 speech perception.

Nevertheless, phonetic training of the AV-HVPT approach seems to not work equally well among learners of different L2 backgrounds (Aliaga-García & Mora, 2009; Hazan et al., 2005). In Hazan et al. (2006), for example, Spanish learners showed significantly higher degrees of sensitivity to labial/labiodental consonants than Japanese learners in audio, visual and audio-visual perception tasks. Furthermore, in the perception of the less visually salient English contrast /l/–/r/, neither Japanese nor Korean learners benefited from the visual codes of the contrasts, although the Korean learners outperformed the Japanese learners in auditory and audio-visual conditions. This was attributed to the lack of distinct/different articulatory gestures of /l/–/r/ in Japanese and Korean (Hazan et al., 2006). To further explain this issue, Hazan et al. (2006) classified speech sounds into three visual categories: a visual category (1) exists both in learners' L1 and L2; (2) exists in the L2 but not in the L1; (3) occurs both in L1 and L2 but is used in different phonetic distinctions in the L1 and L2. According to Hazan et al. (2006), as language learners may lose sensitivity to non-native visual codes, L2 visual categories in conditions (2) and (3) are predicted to be more difficult to learn than those in condition (1).

Despite the wide success of the AV-HVPT approach in laboratories, scarce evidence is available regarding whether this approach can be employed in classrooms. Among the available evidence, Li and Somlak (2019) compared the effects of AV with AO (audio-only) perception training on Mandarin speakers' learning of English interdental–alveolar contrasts (/θ/–/s/ and /ð/–/z/) in a college English course. The participants were exposed to either AV or AO recordings of native English speakers' productions of seven English poems containing the target contrasts. Participants' perception and production of the two contrasts was examined with a pre-, post-, and delayed post-test. Compared with the respective pre-, post-, and delayed post-test, the AV group significantly outperformed the AO group in both the perception and the production testing results. Specifically, mean accuracy improved from about 30% (pre-test) to approximately 60% (post- and delayed post-test), and 1 out of 41 participants received 100% pronunciation accuracy after the teaching programme). However, stimulus variability in Li and Somlak's (2019) training was rather low: only two voices were used and the number of times that the target sounds occurred in the poems was very limited. In addition, the articulatory alveolar/interdental target gestures were fairly visually salient. The present study intends to increase the variability of the phonetic training: more native English speakers for material recording, more stimuli with target sounds in the learning materials, and more learning sessions for the contrasts /s/ vs. /θ/, and /ð/ vs. /z/. In addition, this study also intends to consider a less visibly salient articulatory contrast – /ɪ/ vs. /i/ (close vs. near-close front unrounded vowels) – to explore whether these type contrasts would also benefit from AV-HVPT training. The following research questions were proposed:

- Research question 1: Can an AV-HVPT improve students' pronunciation accuracy of dental fricatives /θ/ and /ð/ in a classroom context?
- Research question 2: Can an AV-HVPT improve students' pronunciation accuracy of the less visually salient sound: near close front unrounded /ɪ/ vowel in a classroom context?

III Methodology

This study was undertaken at a university in northwest China in an intact English course: an already-formed group which did not undergo selection or manipulation for the purpose of the study. A pilot study was conducted first to find target sounds that were difficult for the participating students to pronounce. Then, the students experienced either an AV-HVPT approach or a traditional teacher-instruction approach in a 12-week learning programme. Their pronunciation of the target sounds was tested before and after the learning programme. All the participating students, speakers, and raters were compensated for their participation.

1 Pilot study

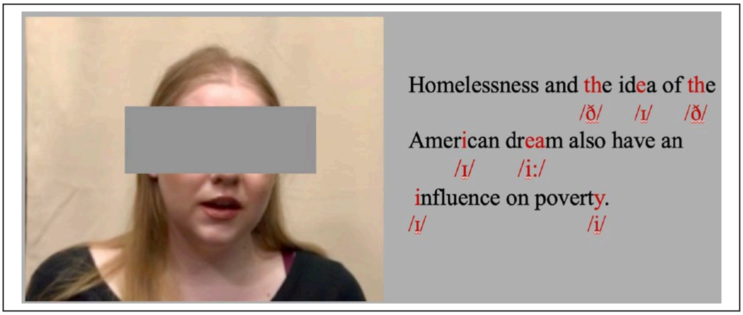
a Participants. Two intact classes of Chinese graduates who majored in medical science volunteered to participate in the study: Class 1: N=36, 16 males, 20 females, mean age = 23.48 years; Class 2: N=34, 15 males, 19 females, mean age = 24.11 years. All of them were Chinese speakers of English who started learning English between 7 and 9 years old. They were taking the *General Academic English* (Lv, 2016) for credit in the spring semester of 2020.

b Selection of target sounds. A pilot study was conducted prior to the main study to determine the sounds with which the participating students had difficulties. It was conducted from Monday to Wednesday in the first week of the semester. Individual students were asked to read aloud the 200-word text ‘The boy who cried wolf’ (Deterding, 2006), which contains all the English vowels and consonants, in a quiet classroom at university. They were given 10 minutes for preparation, during which they were allowed to look up unknown word(s) with an *Oxford English–Chinese Dictionary* next to them (students did not make use of the dictionary). Productions were audio recorded with a Roland-05 recorder (setting: 16-bit mono channel, 44.1 kHz), saved in WAV format, and sent to one male rater and one female rater via Dropbox for assessment. The raters were both phonetically trained native English speakers. They were asked to note and transcribe individual students’ mispronounced sounds. Any difference between the two raters’ assessments was reassessed by a third native English speaker. According to the results, the most severe problems lay in distinguishing /θ/–/s/, /ð/–/z/ and /ɪ/–/i/; more than 94% of the students in both classes replaced /θ/, /ð/ and /ɪ/ with /s/, /z/ and /i/, respectively. Therefore, /θ/–/s/, /ð/–/z/ and /ɪ/–/i/ were selected as the target contrasts for the main study.

2 Main study

a Learning programme. Given that the study was embedded in an intact English course, the learning programme followed its schedule. Specifically, the course lasted for 12 weeks with two 90-minute (with a 10-minute interval for rest) sessions per week, yielding 24 sessions in total. Class 1 experienced an AV-HVPT learning approach, while Class 2 were taught with a traditional teacher instruction by a familiar teacher for them.

b Learning materials. The learning materials of the English course were based on the syllabus of the course, which covered units 5–8 of the textbook *General Academic*



Homelessness and the idea of the
/ð/ /ɪ/ /ð/
American dream also have an
/ɪ/ /iː/
influence on poverty.
/ɪ/ /i/

Figure 1. A screenshot from the audio-visual interface.
Note. Part of the speaker's face was covered to preserve anonymity.

English (Lv, 2016). Each unit included an audio listening exercise, intensive reading of two English texts (Text A and Text B of approximately 800–1300 words each) and a writing task.

c Audio-visual material developed for AV-HVPT interface (Class 1). Four male and four female native English speakers (mean age: 26.8 years old) who were pursuing graduate degrees in linguistics at a university in England were recruited to read the texts, each of them read one text. All the readers were born and raised in southern England. They were told that their readings would be used to teach English pronunciation. They were asked to pronounce the target sounds according to the descriptions of International Alphabetic Symbol chart (International Phonetic Association, 1999). The recordings were conducted in a soundproof booth at the speakers' university with a high-quality video recorder (Lilliput A7S). The audio-visual recordings were edited with a speaker's face shown on the left of the screen, while a synchronized text sentence was displayed on the right of the screen with the target sounds[2] in red (see Figure 1).

d Learning procedure. Both Class 1 and Class 2 began each unit with lead-in activities that followed the tasks on the textbook (e.g. discussions on the topic of the passage, understanding of global structure, answering topic-related questions), then followed by a detailed paragraph-by-paragraph learning of Text A and Text B, during which the phonetic training was carried out in Class 1, while Class 2 followed a teacher-led learning approach.

In the detailed learning of each paragraph of a text, Class 1 was asked to do a mimic-reading task: students were first asked to watch the audio-visual recording of a native English speakers' reading of the paragraph for the first time, then read after (mimic) the speaker's reading sentence by sentence. The mimic reading of each paragraph repeated three times. After this, students were asked to answer questions related to the content of the paragraph, with feedback (correct answers) being provided by the teacher.

Class 2 began with a read-aloud task of a paragraph to themselves. After that, 2~4 students were asked to read the paragraph in front of the whole class; the teacher corrected any incorrectly pronounced sounds by asking the whole class to read after her.

Following this, the teacher then read the paragraph and the students were asked to listen to / watch her reading carefully. After that, the same as Class 1, the students in Class 2 were asked to answer questions related to the content of the paragraph, with feedback (correct answers) being provided by the teacher.

e *Tests.* To test their accuracy in pronouncing /θ/–/s/, /ð/–/z/ and /ɪ/–/i/, the students were asked to perform a read-aloud task before (pre-test) and right after (post-test) the learning programme. The testing materials were 19 randomized English sentences containing each of the target sound in 15 words with the target sounds /θ, s, ð, z/ in different phonetic positions: for word-initial $N=5$, for word-middle $N=5$, and for word-final $N=5$; /ɪ, i/ were only embedded in word-initial ($N=5$, e.g. *eat – it*) and word-middle ($N=10$, e.g. *deep – zip*) positions, yielding to 90 stimulus words in total.

Pre- and post-test were both conducted in a quiet classroom with the same equipment and followed the same procedure as in the pilot study. The students' readings were randomized and sent to the three raters for assessment via Dropbox after post-test was completed. Similar to some prior studies (e.g. Li, 2016; Li & Somlak, 2019), the raters were asked to judge the pronunciation accuracy of the target sounds with a 5-point Likert scale (0=totally wrong; 5=totally correct). Average scores of the three raters' responses were calculated for each stimulus and converted into percentage for statistical analysis. Individual raters' assessment on each stimulus were input into SPSS and analysed for reliability (Cohen's Kappa), which was found to be high among the three raters (.93).

IV Results

The two groups were comparable before the learning programme was carried out. A one-way ANOVA test was performed with Class (Class 1 or Class 2) coded as the between-group factor and individual students' pronunciation accuracy of the three sounds (in percentage) coded as the dependent variables. As seen in Table 1, the groups

Table 1. Descriptive data on the students' pronunciation of /θ, ð, ɪ/ in pre- and post-test.

Target sound	Test	Class	Mean (%)	Standard error
/θ/	Pre-test	Class 1	37.0	2.1
		Class 2	43.7	3.2
	Post-test	Class 1	76.8	4.0
		Class 2	44.1	3.3
/ð/	Pre-test	Class 1	33.8	2.2
		Class 2	41.0	3.3
	Post-test	Class 1	75.5	4.2
		Class 2	40.0	3.3
/ɪ/	Pre-test	Class 1	41.6	3.3
		Class 2	40.5	3.2
	Post-test	Class 1	40.3	3.4
		Class 2	43.0	3.6

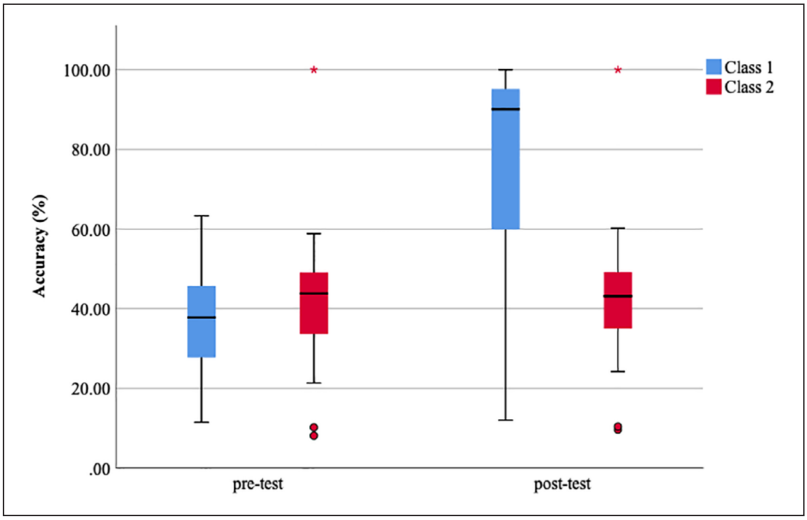


Figure 2. Pronunciation accuracy of /θ/ by Class 1 and Class 2 in pre-test and post-test.

did not perform significantly differently in pronouncing /θ/ (F(1, 69) = 3.09, $p > .05$), /ð/ (F(1, 69) = 3.45, $p > .05$). $p > .05$) or /ɪ/ (F(1, 69) = 0.06, $p > .05$) in pre-test. In post-test, Class 1 displayed significantly higher accuracies than Class 2 in pronouncing /θ/ (F(1, 69) = 37.80, $p < .001$) and /ð/ (F(1, 69) = 43.61, $p > .001$) but not in pronouncing /ɪ/ (F(1, 69) = 4.16, $p > .05$, $p = .06$), which remained in rather low accuracy scores (approximately 40%) in both the pre- and post-test.

One-way ANOVA tests were conducted to examine whether the two classes had a significant pronunciation improvement from the pre-test to post-test. Individual students' accuracies in pronouncing /ð, θ, ɪ/ were coded as the dependent variables, and test (pre- or post-test) was coded as the independent factor. The results showed that Class 1 significantly improved /θ/ (F(1, 71) = 51.10, $p < .001$) and /ð/ (F(1, 71) = 61.80, $p < .001$) in post-test. Their performance in the pronunciation of /ɪ/, however, was non-significantly different between the pre- and post-tests (F(1, 71) = 0.82, $p = 0.78$). Class 2 displayed no significant differences between the pre-test and post-test in pronouncing /θ/ (F(1, 67) = 0.01, $p = .93$), /ð/ (F(1, 67) = 0.05, $p = .83$) or /ɪ/ (F(1, 67) = 0.02, $p = .89$).

Results revealed individual variability, some students scoring 100% both in post-test for the three sounds, for example (see Figures 2–4). It could also be observed that Class 1's post-test exhibited a wide range of responses, with some students reaching 100% correct responses for /θ/ and /ð/, while others remained below 20%. Additional ANOVA tests were performed with individual students' pronunciation accuracy of the target sounds in pre- and post-test as the dependent variables, while their age, gender, years of English study, level of English proficiency were coded as the between-participants factors. None of these factors were found to have a significant effect on the pronunciation performance of the participants in any of the tests ($p > .005$).

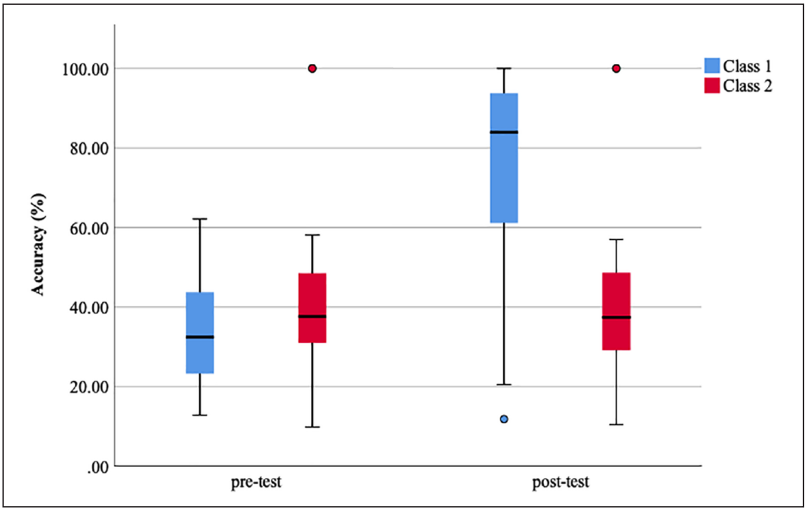


Figure 3. Pronunciation accuracy of /ð/ by Class 1 and Class 2 in pre-test and post-test.

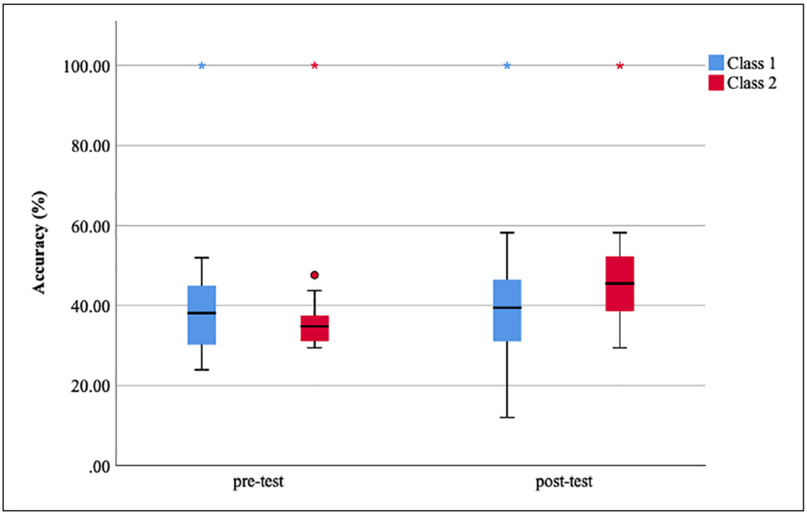


Figure 4. Pronunciation accuracy of /ɪ/ by Class 1 and Class 2 in pre-test and post-test.

V Discussion

This study revealed that the group having received AV-HVPT showed a significant improvement between the pre- and post-test in the pronunciation of /θ, ð/, whereas the group having received teacher training did not change significantly between the two tests. Considering that Class 1 was able to observe the articulatory gestures of the target sounds during the AV-HVPT learning programme while Class 2 was not exposed to such audio-visual interface of the target sounds' articulatory information, this finding may confirm the effectiveness of using the AV-HVPT approach in L2 sound learning in classroom settings. It is in line with findings in Li and Somlak (2019), in which audio-visual training also led to the participants' significant improvement in the accurate pronunciation of /θ, ð/. Nevertheless, in the present study, the AV recording used for Class 1's learning not only included the observation of articulatory gestures but also involved visual phonetic symbols of the target sounds. According to the educational system in China, as a necessary skill, all Chinese learners of English learn how to read and spell English phonetic symbols since middle schools. Class 1's pronunciation improvement, therefore, might have benefited from having the phonetic symbols provided in the AV recordings.

In addition, as discussed in the literature review, in Li and Somlak (2019) the participants' average pronunciation improvement was approximately 30% in pronouncing both /θ/ and /ð/. The present study presented a larger degree of pronunciation improvement than previous studies after similar AV-HVPT training methodology had been implemented as in Li and Somlak (2019). The relatively larger degree of pronunciation improvement found in this study might be explained by its supplementary methodology, which aimed at increasing the variability in the HVPT approach (Barriuso & Hayes-Harb, 2018). Specifically, students in Class 1 were exposed to a larger number of stimuli produced by more native speakers as well as more trials as part of the learning programme. Such increased variability may have also enhanced variability may have also enhanced learners' performance in the present study. In addition, the learning program in this study was longer than that of Li and Somlak (2019), this fact may have also enhanced L2 sound learning in the present study.

Interestingly, the teacher-led group did not undergo significant pronunciation improvement, unlike studies which show pronunciation improvement of this type of training (Neri et al., 2008). While one possible reason considered is lack of accurate input on the part of the teacher (Li & Somlak, 2019), our data seem to suggest that audio-visual aids facilitated the pronunciation of particularly difficult sounds before a teacher-led pronunciation class, which may need more time to attract the attention of the learners towards such articulatory gestures. Moreover, Class 1 experienced a classroom integrated procedure in addition to the AV-HVPT approach. Their relatively better performance in relation to Class 2, therefore, may not be fully attributed to the AV-HVPT learning approach.

Another finding was Class 1's post-test performance in pronouncing /ɪ/, having been considered to be a less articulatory salient sound than /θ, ð/. Despite the fact that it was embedded in the same AV-HVPT programme and targeted in the teacher-led group, neither group underwent a change in their pronunciation performance towards a closer and less fronted sound than /i/. This finding is in line with findings in Hazan et al. (2006), in which L2 learners of English received much lower scores in visual conditions for the less visually salient /l/–/r/ contrast than the highly salient labial/labiodental contrast. In the same vein, the differences in pronunciation performance of Class 1 after the AV-HVPT between the pronunciation of /θ, ð/ and /ɪ/ might be explained by the fact that the articulatory information of interdentals /θ, ð/ is more salient when compared with alveolars /s, z/ (tongue visibility) and thus easier to observe than that of the near-close front unrounded vowel /ɪ/, versus a closer /i/ articulatory configuration, which could be a slight difference in lip spread configuration.

It should be mentioned that some individual differences were found in the pronunciation tests. Six out of 36 students in Class 1 scored 100% in pronouncing /θ/ and /ð/ in the post-test, while another student received 100% in pronouncing /θ/ and 94.1% in pronouncing /ð/. None of the seven students' scores were above 60% in the pre-test. In comparison, eight of the students' scores in the post-test either hardly increased or slightly decreased in pronouncing /θ/ and/or /ð/, showing no changes after pronunciation instruction (Pennington & Rogerson-Revell, 2019). Here contextual, motivational, affective or strategic factors may be intervening (Baran-Łucarz, 2012; Szyszka, 2015; Trofimovich et al., 2015). For example, in informal discussions after the study was completed, those who scored 100% in pronouncing the target sounds in both the pre-test and the post-test all indicated that they were enthusiastic about English learning, and they liked watching English movies / TV series and listening to English songs in their spare time. This may have formed a strong motivation for English learning and facilitated their English learning. Moreover, although all the participating students had been learning English as an L2 for 9~13 years since childhood, none of them had any experience staying in an English-speaking community for more than a week; only four students (one in Class 1; three in Class 2) indicated that they had travelled shortly abroad; another five students (four in Class 1; one in Class 2) reported that they had been taught by a native-English teacher for half an academic term (about 12 lessons). The analysis of individual factors will be addressed in future exploration of pronunciation performance in training studies.

VI Conclusions

This study examined the effects of an AV-HVPT approach on L2 sound learning in a classroom setting. Compared with the teacher-led taught students, those who followed the AV-HVPT approach had a significant improvement in the pronunciation of English /θ, ð/, whereas neither class showed a significant change in pronouncing English /ɪ/. This finding indicated that less salient articulatory information benefits significantly less from AV-HVPT procedures. Accordingly, although phonetic training following an AV approach has received wide success in laboratory settings (e.g. Hazan et al., 2005; Li, 2016), it might only be effective when the articulatory gestures are saliently observable. Moreover, it was found that increased variability in the size of training stimuli and number of native speakers of the AV-HVPT protocol group produced higher gains than in previous studies with similar AV-HVPT procedures (Li & Somlak, 2019). We speculate that for L2 sound learning, the HVPT approach is as effective in classroom settings as in laboratory settings.

The findings of this study suggest some pedagogical implications. For example, for learning L2 sounds with saliently observable articulatory gestures (e.g. interdentals, bilabials), audio-visual materials produced by multiple native speakers of the L2 that involve a large number of exemplars of target sounds can be more extensively used in classrooms, which can help learners observe the articulatory information of the sounds. For the learning of L2 sounds with less saliently observable (or unobservable) articulatory gestures (e.g. velars, uvulars, back vowels), as shown in the present study, using the AV-HVPT approach may not be sufficient. Instructors may need to adopt techniques to help learners interpret non-visible articulatory gestures, most likely with the use of technology with apps, virtual reality simulators, etc., which can nowadays be found on specific pronunciation websites.

This study had a number of limitations. For example, the students' pronunciation accuracy was assessed auditorily by native English speakers. Although the interrater reliability was found to be high, no acoustic analysis was performed to further investigate more detail in their pronunciation performance. Due to practical reasons, a delayed post-test was not conducted in this study. It was therefore unclear whether and how long the positive learning effects of AV-HVPT could last. These issues will be taken into consideration in future studies.

Acknowledgements

This article and the research behind it would not have been possible without the exceptional support of the participating students, native English speakers and raters. I am also grateful for the insightful comments offered by the anonymous peer reviewers. Their generosity and expertise have improved this study innumerable ways; those inevitably remain are entirely my own responsibility.

Compliance with Ethical Standards

This study compliance with ethical standards. All the participants were adults. They signed a consent form and volunteered to participate in the study.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Ying Li https://orcid.org/0000-0003-1783-9083

Notes

1. 'Incorrectly pronounce' in this article refers to non-native-like realization of L2-English speech sounds.
2. Letters and corresponding phonetic symbols: according to the educational system in China, as a necessary skill, all Chinese learners of English learned how to read and spell English phonetic symbols since middle school, though it is usually not examined in the final exams.

References

Aliaga-García, C., & Mora, J.C. (2009). Assessing the effects of phonetic training on L2 sound perception and production. In Watkins, M.A., Rauber, A.S., & B.O. Baptista (Eds.), *Recent research in second language phonetics/phonology: Perception and production* (pp. 2–31). Cambridge Scholars.

Bada, E. (2001). Native language influence on the production of English sounds by Japanese learners. *The Reading Matrix*, *1*, 1–15.

Baran-Łucarz, M. (2012). Individual learner differences and accuracy in foreign language pronunciation. In Pawlak, M. (Ed.), *New perspectives on individual differences in language learning and teaching* (pp. 289–303). Springer.

Barriuso, T.A., & Hayes-Harb, R. (2018). High variability phonetic training as a bridge from research to practice. *CATESOL Journal*, *30*, 177–194.

Bernstein, L.E., Auer, E.T., Jr., Jiang, J., & Eberhardt, S.P. (2013). Auditory perceptual learning for speech perception can be enhanced by audiovisual training. *Frontiers in Neuroscience*, *7*, 34.

Best, C.T., & Strange, W. (1992). Effects of phonological and phonetic factors on cross-language perception of approximants. *Journal of Phonetics*, *20*, 305–330.

Best, C.T., & Tyler, M. (2007). Nonnative and second-language speech perception. In Bohn, O.-S., & M.J. Munro (Eds.), *Language experience in second language speech learning: In honour of James Emil Flege* (pp. 13–34). John Benjamins.

Bradlow, A.R., Pisoni, D.B., Akahane-Yamada, R., & Tohkura, Y.I. (1997). Training Japanese listeners to identify English /r/ and /l/: IV, Some effects of perceptual learning on speech production. *The Journal of the Acoustical Society of America*, *101*, 2299–2310.

Carlet, A. (2017). L2 perception and production of English consonants and vowels by Catalan speakers: The effects of attention and training task in a cross-training study. Unpublished doctoral dissertation, Universitat Autònoma de Barcelona, Barcelona, Spain.

Chen, T. (2001). Audiovisual speech processing. *IEEE Signal Processing Magazine*, *18*, 9–21.

Deterding, D. (2006). The North Wind versus a Wolf: Short texts for the description and measurement of English pronunciation. *Journal of the International Phonetic Association*, *36*, 187–196.

Flege, J.E. (1995a). Second language speech learning: Theory, findings, and problems. *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, *92*, 233–277.

Flege, J.E. (1995b). Two procedures for training a novel second language phonetic contrast. *Applied Psycholinguistics*, *16*, 425–442.

Flege, J.E. (2003). Assessing constraints on second-language segmental production and perception. In Schiller, N.O., & A.S. Meyer (Eds.), *Phonetics and phonology in language comprehension and production, differences and similarities* (pp. 319–355). Mouton de Gruyter.

Fowler, C.A. (1994a). Speech perception: Direct realist theory. In Asher, R.E., & M.Y. James (Eds.), *The encyclopedia of language and linguistics* (pp. 4199–4203). Pergamon.

Fowler, C.A. (1994b). Invariants, specifiers, cues: An investigation of locus equations as information for place of articulation. *Perception and Psychophysics*, *55*, 597–610.

Fowler, C.A. (1996). Listeners do hear sounds, not tongues. *The Journal of the Acoustical Society of America*, *99*, 1730–1741.

Hazan, V., Sennema, A., Faulkner, A., et al. (2006). The use of visual cues in the perception of non-native consonant contrasts. *The Journal of the Acoustical Society of America*, *119*, 1740–1751.

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*, 360–378.

Hirata, Y., & Kelly, S.D. (2010). Effects of lips and hands on auditory learning of second-language speech sounds. *Language and Hearing Research*, *53*, 298–310.

International Phonetic Association. (1999). *Handbook of the international phonetic association: A guide to the use of the international phonetic alphabet*. Cambridge University Press.

Iverson, P., Hazan, V., & Bannister, K. (2005). Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English /r/–/l/ to Japanese adults. *The Journal of the Acoustical Society of America*, *118*, 3267–3278.

Kachru, B.B. (1990). World Englishes and applied linguistics. *World Englishes*, *9*, 3–20.

Krashen, S.D. (1981). *Second language acquisition and second language learning*. University of Southern California.

Krashen, S.D. (1985). *Inquiries and insights: Second language teaching: Immersion and bilingual education, Literacy*. Alemany Press.

Levis, J. (2020). Revisiting the intelligibility and nativeness principles. *Journal of Second Language Pronunciation*, *6*, 310–328.

Li, Y. (2016). Audiovisual training effects on L2 speech perception and production. *International Journal of English Language Teaching*, *3*, 14–36.

Li, Y., & Somlak, T. (2019). The effects of articulatory gestures on L2 pronunciation learning: A classroom-based study. *Language Teaching Research*, *23*, 352–371.

Liberman, A.M., Cooper, F.S., Shankweiler, D.P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*, 431–461.

Liberman, A.M., & Mattingly, I.G. (1985). The motor theory of speech perception revised. *Cognition*, *21*, 1–36.

Liberman, A.M., & Whalen, D.H. (2000). On the relation of speech to language. *Trends in Cognitive Sciences*, *4*, 187–196.

Lively, S.E., Logan, J.S., & Pisoni, D.B. (1993). Training Japanese listeners to identify English /r/ and /l/: II, The role of phonetic environment and talker variability in learning new perceptual categories. *The Journal of the Acoustical Society of America*, *94*, 1242–1255.

Logan, J.S., Lively, S.E., & Pisoni, D.B. (1991). Training Japanese listeners to identify English /r/ and /1/. *The Journal of the Acoustical Society of America*, *89*, 874–886.

Logan, J.S., Lively, S.E., & Pisoni, D.B. (1993). Training listeners to perceive novel phonetic categories: How do we know what is learned? *The Journal of the Acoustical Society of America*, *94*, 1148–1151.

Lv, Y.B. (2016). *General academic English*. Higher Education Press.

Neri, A., Mich, O., Gerosa, M., & Giuliani, D. (2008). The effectiveness of computer assisted pronunciation training for foreign language learning by children. *Computer Assisted Language Learning*, *21*, 393–408.

Ortega-Llebaria, M., Faulkner, A., & Hazan, V. (2001). Auditory-visual L2 speech perception: Effects of visual cues and acoustic-phonetic context for Spanish learners of English. In Massaro, D.W., Light, J., & K. Geraci (Eds.), *AVSP 2001 – International Conference on Auditory-Visual Speech Processing*. ISCA.

Pennington, M.C., & Rogerson-Revell, P. (2019). *English pronunciation teaching and research* (*Vol. 10*, pp. 978–988). Palgrave Macmillan.

Picard, M. (2002). The differential substitution of English/θ ð/in French: The case against underspecification in L2 phonology. *Lingvisticæ Investigationes*, *25*, 87–96.

Pisoni, D.B., & Lively, S.E. (1995). Variability and invariance in speech perception: A new look at some old problems in perceptual learning. In Strange, W. (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 429–455). York Press.

Ross, L.A., Saint-Amour, D., Leavitt, V.M., Javitt, D.C., & Foxe, J.J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex*, *17*, 1147–1153.

Sennema, A., Hazan, V., & Faulkner, A. (2003). The role of visual cues in L2 consonant perception. In Solé, M.J., Recasens, D., & J. Romero, (Eds.), *Fifteenth International Congress of Phonetic Sciences*. Casual Productions, pp. 135–138.

Sumby, W.H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America*, *26*, 212–215.

Szyszka, M. (2015). Good English pronunciation users and their pronunciation learning strategies. *Research in Language*, *13*, 93–106.

Trofimovich, P., Kennedy, S., & Ann Foote, J. (2015). Variables affecting L2 pronunciation development. In Reed, M., & J.M. Levis (Eds.), *The handbook of English pronunciation* (pp. 353–373). Wiley Blackwell.

Wang, Y., Behane, D., & Jiang, H. (2008). Effects of training modality on audio-visual perception of nonnative speech contrasts. *Canadian Acoustics*, *36*, 120–121.