International Journal of English Language Education
ISSN 2325-0887
2023, Vol. 11, No. 1

Investigating the Efficacy of Using Online Resources for Production Training in Learning Non-Native Vowel Contrasts

Ajwaad Aljohani

English Language Institute, King Abdulaziz University, P.O. Box 42801, Zip Code 21551, Jeddah, Saudi Arabia

Wafaa Alshangiti (Corresponding author)

English Language Institute, King Abdulaziz University, P.O. Box 42801, Zip Code 21551, Jeddah, Saudi Arabia

Received: December 8, 2022 Accepted: December 22, 2022 Published: December 24, 2022

doi:10.5296/ijele.v11i1.20601 URL: https://doi.org/10.5296/ijele.v11i1.20601

Abstract

The current study investigates the benefits of virtual phonetic training on second language (L2) vowel production and examines whether production training improves perception. While the benefits of production instructions were investigated with instructions that were delivered by either a virtual or a face-to face teacher, a few studies investigated the benefits of online resources for second language (L2) learning with tutor audio instructions and feedback. Fifty-five Arabic speakers (21 served as a control group) took part in 3 one-to-one training sessions and were trained on producing seven Standard Southern British English (SSBE) vowels (/ ɪ, ɛ, ʌ, ɑː, ɔː, ɒ, ʊ/) that have been found to be difficult for Arabic learners of English. Improvements were assessed by their vowel identification, category discrimination accuracy and their vowel intelligibility. The results showed some improvements in vowel intelligibility and that vowel learning paved the way for improvement in perception, relative to the control group. The results indicate that using web-based tools along with audio instructions are beneficial for L2 learners and confirmed previous findings about the relationship between improvement in speech perception and production.

Keywords: pronunciation training, online ultrasound videos, L2 speech perception and production

1. Introduction

Adult learners are often found to have difficulties in acquiring speech sounds of L2. This difficulty is usually attributed to, among other factors, the relationship between learners' first language (L1) phonetic system and that of the L2. Some theoretical models (Best & Tyler, 2007; SLM, Flege, 1995; SLM-r, Flege & Bohn, 2021) suggest that learning speech in L2 was shaped by perceptual biases induced by the L1 phonetic system. For example, the Perceptual Assimilation Model PAM (Best, 1995; Best & Tyler, 2007) describes the difficulty that L2 learners face as a result of assimilating L2 contrasts into similar L1 phonetic categories. For example, Greek learners of English assimilate the English tense and lax vowels /i:/- /ɪ/ to their L1 vowel /i/ (Lengeris, 2009). The bias in L2 perception that stems from the interference between learners' L1 and L2 phonetic systems, can also cause some production difficulties. For instance, Italian learners of English face difficulties with producing the vowel contrast /ɑ/-/ʌ/ (Flege et al., 1999), whereas Catalan learners of English find difficulties in producing /i/-/ ɪ/ contrast (Cebrian, 2007).

Despite such difficulty, explicit phonetic instructions were found to help L2 learners improve their L2 phonemes production (Camus, 2019) and perception (Kartushina et al., 2015). In order for L2 learners to distinguish and produce L2 phonemes differently from those of L1, learners need to receive explicit instructions, and extensive practice. For example, when L2 learners acquire a target language, pronunciation becomes one of the challenges due to their L1 knowledge besides other factors, such as age and exposure (Flege, 1995; MacKay, 2006). However, learning how to pronounce L2 sounds requires not only an abstract knowledge of sounds, but also the articulators that are involved in producing certain L2 speech sounds, i.e., learners need to see some visuals of moving and still articulators and get some instructions and feedback that are tailored to each individual learner (McCrocklin, 2016). This is especially important for L2 learners as one of the main requirements for their proficiency in the target language is the level of their intelligibility which is reflected by their pronunciation. As a result, a number of pronunciation training practitioners train L2 learners to be more intelligible, rather than having native-like pronunciation (cf. Levis, 2005; Alshangiti & Evans, 2014; Thomson and Derwing, 2015; Gilakjani, 2012).

Indeed, Pronunciation training has been conducted using different techniques that involve computer assisted pronunciation training (CAPT) where the instructions and feedback were delivered by either a face-to-face tutor (Hattori, 2010; Alshangiti & Evans, 2014), or by a virtual tutor (e.g., Bälter et al., 2005; Engwal & Bälter, 2007; Engwal, 2012). Moreover, a number of studies used automatic speech recognition (ASR) based on computer assisted language learning (CALL) to train L2 learners on pronunciation (e.g., Dalby and Kewley-Port, 1999; Chou, 2005; Neri et al., 2008; McCrocklin, 2016). In fact, Neri et al (2008) developed an ASR-based on the CAPT system for training Dutch L2 speakers in pronouncing eleven Dutch phonemes that have been found to be problematic for L2 speakers who have different L1 background. Participants were asked to pronounce minimal pairs of each phonemic contrast, and if they pronounced the word correctly, they were rewarded with positive feedback, in the form of a smiley face along with the orthographic representation of the utterance. However, if the mispronounced words were detected by the ASR algorithm, the participants were prompted

to reproduce the word, with a disappointed face along with the orthographic representation of the utterance which was displayed on the screen with the corresponding letter(s) coloured in red. The feedback helped L2 learners to improve their pronunciation, though the difference was not significant compared to the control group. By the same token, other studies combined CAPT systems with phonetic-based approaches by creating a virtual language teacher as a vowel-learning tool for L2 learners from different L1 backgrounds (e.g., Wik, 2011), and found that over two training sessions, there was some improvement in vowel production compared to the first session. These studies showed that the CAPT system has a number of advantages (e.g., no need for teachers, immediate feedback). That said, the above-mentioned studies did not find a great improvement as one might expect. This is possibly because the feedback is already programmed, and what one learner might find beneficial may not be generalised to all learners. Another reason, for Neri et al. (2008) and Wik (2011) studies, is that they trained L2 learners from different L1 backgrounds, and therefore their difficulty of perceiving or producing L2 phonemes might differ according to the relationship between L1 & L2 phonetic systems, and thus need explicit individual feedback that suits their needs. More positive results, however, were found when training was focussed on learners from the same L1 background (Hattori, 2010; Alshangiti & Evans, 2014; Kartushina et al., 2016; Kartushina & Martin, 2019; Feng, 2020), using some visual aids (spectrograms or visual display of animated midsagittal section or vowel formants), with explicit feedback.

The feedback in such training studies can be indirect, displaying visual information that informs articulations, e.g., using spectrogram or wave forms to make comparisons between the target speech and the L2 learners' production (e.g., Akahane-Yamada et al., 1998; Carey, 2004; Hattori, 2010; Olson, 2014), or direct, displaying articulatory processes of learners (e.g., Pillot-Loiseau et al., 2013). For example, direct feedback methods were used in a number of studies that trained pronunciation using visual ultrasound images which were found to improve L2 learning. This system helps learners to visualise the tongue movements of the target phonemes (e.g., Wilson et al., 2006; Gick et al., 2008; Abdel et al., 2015; Mozaffari et al., 2019). Moreover, they are safe, portable, and relatively inexpensive to use. These advantages encouraged researchers to use this system to show the subtle movements of the tongue that are important for producing certain speech sounds. It also enables learners to visualize the front and the back of the tongue simultaneously. When, for instance, Japanese learners of English learn the difference between /r/-/l/, they can see the important part of the tongue that is responsible for producing either consonant, depending on its position in the syllable, i.e., onset or coda (Campbell, 2004). They can also see the tongue root, tongue dorsum, and the tongue tip. Ultrasound techniques were also found to be beneficial for improving L2 vowel production (Mackay, 1977; d'Apolito et al., 2017). For example, Pillot-Loiseau et al. (2013) used ultrasound images to show direct feedback of Japanese speakers learning French /u/ and /y/ vowels and found, after three training sessions, that learners improved their production of the French vowels /u/ and /y/. Similarly, d'Apolito et al., (2017) trained Italian speakers in the American English contrasting speech sounds /ɑ/-/ʌ/ over a one-hour session using real-time ultrasound feedback of learners' production and found that their production had improved after training.

In spite of the advantages of ultrasound imaging for improving production, using it in a pedagogical set for teaching L2 can be expensive, especially when fully equipped laboratories or practitioner groups are not available in some institutions. One easier and cheaper alternative, therefore, can be using ultrasound videos as visual aids to teach L2 learners how to produce certain phonemes following the native speakers' production. This method would need a tutor to teach L2 learners how to read the information and where to focus their attention, that is, the front or back of the tongue for instance. Given that teaching L2 speakers difficult phoneme contrasts can be challenging in a classroom setting, virtual sessions can be offered to those who need them. So, the teachers can use online resources to teach L2 phonemic contrast and use reliable resources that show clear visuals of the articulators.

The current study used one of the online resources that provides a midsagittal view of ultrasound imaging with lip videos proceeding the tip of the tongue. This was particularly important as this source makes it easier to show the lip which guides the learners visually to the tongue and the lip movements when a word is produced. Learners who have no experience with ultrasound, might find it difficult to follow, localize or interpret the tongue movements (cf. Mozaffari et al., 2019). Therefore, in the current study, a tutor gave instructions and feedback along with a teaching tool that was chosen from various online resources and applications that were designed for pronunciation training.

With the development of different technologies that implement CAPT in L2 learning, and with the increasing need for online tools and resources, a growing number of pronunciation training applications and websites have offered pronunciation learning platforms. For example, one of the best-known websites for pronunciation instructions is the flash animated project of the university of Iowa (https://soundsofspeech.uiowa.edu), which has a related application; the Iowa sound of speech application (Bangun & Liontas, 2019), that was found to be beneficial in teaching American English vowels and consonants. Moreover, some websites such as English Central (https://www.englishcentral.com/browse/videos), and Voice Tube (https://www.voicetube.com/) are decent resources for shadowing and imitating, while others allow learners to record themselves, practice pronunciation and even share their pronunciation by email or social media such as Fotobabble website (http://www.fotobabble.com), (cf. Yoshida, 2018 for review). However, most online resources might not accurately judge the pronunciation, or merely give instructions without feedback, and even if some provide feedback, it might not be suitable for individuals (cf. Derwing & Munro, 2015). However, some other applications, with attractive and high-end technology, are mainly commercial rather than useful for pedagogy, while others are specifically for certain varieties of English (Kaiser, 2017).

Other studies (e.g., Alshangiti et al., 2019) used customised low-tech videos and animated images of the midsagittal section of the tongue to illustrate vowel articulation with a face-to face tutor. In this study,they found little improvement of vowel production after five training sessions. One possibility is that the participants had excessive visual cues as they were presented with (videos of native speakers, the animated midsagittal section, and the tutor) which might be a lot to process in five sessions. One way the current study could avoid the richness of visual cues, is to have the instructions in audio form. Using audio instructions and feedback from a tutor alongside the online tools means that there are no extra visual cues that

might distract learners from the tools they are presented with. Instead, they can follow the visual display while receiving audio instructions and feedback. This method was chosen to enable learners to ask and get feedback that suits them, along with visualising Ultrasound Tongue Imaging (UTI) films with lip videos for single word production.

The current study, thus, combines the use of online resources, that are created as a product of researchers' collaborations at some of the best Scottish universities (Lawson, et al., 2018), with audio instructions and feedback from a tutor to train Saudi learners of English on vowel pronunciation. For this study, the SSBE accent of these words (7 were chosen to match the investigated vowels) was chosen from the 'Dynamic Dialect' entry from 'Seeing Speech' website (https://seeingspeech.ac.uk/), as Arabic speakers have been known to find some of the SSBE vowels problematic (Evans & Alshangiti, 2018). This difficulty might be related to the difference between the rich vowel inventory for SSBE (20 vowels; cf. Carley et al., 2017) compared to the small Arabic vowel inventory (6 vowels that are contrasted by duration, Al-ani, 2014), which makes mapping L2 vowels onto L1 categories challenging unlike the consonants (cf. Evans & Alshangiti, 2018). By training Arabic learners of English in pronouncing these vowels, the current study investigates the effectiveness of using online resources as visual aids with audio instructions and feedback from an instructor. This was used to investigate whether learners can improve their pronunciation after applying this method of training, and whether this improvement proceeds to improvement in vowel perception.

Training in one speech domain might proceed to another. Some studies show that training in perception proceeds to improving production (e.g., Thomson, 2011; Kartushina et al., 2015; Shinohara & Iverson, 2018; Lengeris, 2018), whereas other studies give mixed results. These studies suggest that training is domain specific, and that production improves production but not perception (e.g., Hattori, 2010; Alshangiti & Evans, 2014). Given the fact that in the current study learners are receiving a great amount of audio instructions and feedback, it might be plausible that their perception might improve even if their attention is focussed on articulation.

2. Method

An experimental research design followed by pre- and post-tests was chosen to determine the effectiveness of training. The validity was ensured by assessing improvements in English vowel production before and after training using pre- and post-tests. The reliability was determined with inferential statistics computed with R software.

2.1 Participants

A total of 55 participants aged between 18-22 years old (median 19 years old) were assigned to two groups: a training (N=34) and a control group (N=21). Participants started learning English when they were 7-18 years old (median 12 years old) with no experience of living in an English-speaking country. A cluster random sampling strategy was selected to have a representation regarding the English language proficiency levels; lower level (A2) and higher level (B2) as shown in table 1.

Table 1. Demographic information of the Saudi participants

Group	Proficiency Level	N	Total
Training group	B2	20	34
	A2	14	
Control group	B2	18	21
	A2	3	

All the participants volunteered to take part in the study and reported no hearing or speech impairment. Participants gave informed consent and had the right to withdraw at any point without explaination.

In addition, five SSBE speakers (2 male, 3 female) aged between 19-57 years old (median 41 years old) identified the words that were produced by the Arabic participants to evaluate any improvements of vowel production after training.

2.2 Apparatus

The pre- and post-tests were conducted via Black Board sessions. Stimuli were played via a laptop with digital output an on-board audio sound card. The same laptop was used to collect the responses via an experimental interface Praat (Boersma &Weenink, 2022). Recordings were made using a digital audio recorder (Zoom H2 Handy Recorder, digital stereo, or 4-channel audio option) at 44.1 kHz, 16-bit resolution.

The articulatory training was completed with an instructor (the second author) with the aid of a web-based interactive articulatory accent data base (Lawson, et al., 2018). Each session took 45 minutes via Black Board sessions.

2.3 Stimuli

2.3.1 Pre- and Post-Tests

The recordings were produced by four native SSBE speakers (2 male, 2 female). The stimuli consisted of seven British English vowels tokens (/ ɑː, ɛ, ɪ, ɔː, ʌ, ɒ, ʊ/) in a /h/-v-/d/ context: *hard, head, hid, hoard, hod, hood*, *hud*. These vowels were chosen for the study as it has been found that they are difficult for Arabic learners (Evans & Alshangiti, 2018). The stimuli were a subset recording from a previous study (Alshangiti et al., 2019) and not used in the training to measure the generalization of new words and speakers.

2.3.2 Training

The stimuli for the training were videos downloaded from a web-based interactive articulatory accent data base (Lawson, et al., 2018) that show ultrasound video of vowel production embedded in common words (e.g., *dress*), along with a side video of the speaker's lips to show

lip rounding or spreading. The SSBE dialect was chosen for the current study, which can be found in this link under London dialect: https://www.dynamicdialects.ac.uk/. Only videos that present the seven vowels being investigated were included in the training.

2.4 Procedure

2.4.1 Pre- and Post-Tests

The pre- and post-tests were conducted via Black Board.

2.4.1.1 Vowel Identification Task

Participants listened to natural recordings of the /h/-V-/d/ words and gave their responses in closed-set options (7 words on one screen as response options). Each response button had a /h/-V-/d/ word written on it (e.g., *hard*) along with more frequent word (e.g., *card*), and each response button was numbered. To give their response, the participants said the number of the response button they think they heard, and the experimenter clicked on the button to move to the next stimuli. This procedure was followed as the session was online, and the experimenter interacted with the participants via Blackboard sessions using the screen share feature. This procedure also ensured that participants could not click buttons without listening to the stimuli. Prior to the task, there was a familiarization session where the experimenter explained to the participants what is expected and how they should respond to each stimulus. They were shown the stimuli and had several trials before starting the task. Participants were able to ask questions and let the experimenter know if the stimuli are clearly audible. They could also adjust the volume to their comfort level. The stimuli were repeated twice: 2 times for each speaker, given a total of 56 stimuli (7 words × 4 speakers × 2 repetitions).

2.4.1.2 Category Discrimination Task

Participants listened to three /h/-V-/d/ words produced by three different speakers in each trial. For each trial, one word was different and two were the same. The participants were asked to judge which one was different (i.e., an oddity task). There was no repetition of the stimuli and there was no feedback. Participants were presented with three response buttons, where the first stimulus was named ‘A’, the second ‘B’, and the third ‘C’, and they told the experimenter which one they thought was the different stimulus. Similar to the identification task, this was online. The stimuli consisted of 4 pairs of words (e.g., *hid-head*), with each pair played six times, three times with *hid* and the other three with *head*, where the odd stimulus played first, second or third. The vowel pairs included were: /ɪ/-/e/, /ʊ/-/ɒ/, /ɒ/-/ʌ/, /ɑː/-/ɔː/ as these sounds are some of the vowel pairs that Arabic learners of English found confusing (Evans & Alshangiti, 2018).

2.4.1.3 Word Recordings

Participants were asked to record the same words they identified in thise presented, i.e., /h/-V-/d/ words. They were presented with power point slides, where they can see one word per slide, to avoid list intonation, and were asked to read the words. There were three repetitions of each word, and the best recording was chosen for vowel intelligibility task to measure any improvement in vowel production. There were 2 recordings, from the pre- and post-tests, which were given to five British English listeners to identify from a closed set, same as the vowel

identification task. There was an additional button labelled 'none' if the listeners thought that none of the words that were produced by our participants matched the response buttons.

2.4.2 Training

There were three one-to-one training sessions of articulatory training with visual information and audio feedback, and an initial practice session for familiarization. Participants in the training group were trained individually to produce seven English phonemes (/ ɑː, ɛ, ɪ, ɔː, ʌ, ɒ, ʊ/) over three training sessions, and each session lasted around 45 minutes. The training sessions were administered on different days over one week depending on participants' availability. The pronunciation training sessions were delivered by an instructor (the second author), an Arabic speaker who is fluent in English and has trained L2 learners in pronunciation. The instructor presented audio-visual materials from Seeing Speech website (Lawson et al., 2018). This website was chosen because it supports the audio-visual learning environment and has audio-articulatory ultrasound for speech production training based on the position of tongue, along with a side view video of the lips which were positioned with the ultrasound video and played simultaneously. This allowed learners to see the tongue movement during vowel production and the lip rounding or spreading at the same time, which makes it more practical than a number of other online resources. In addition, this website offers downloadable videos of the ultrasounds, so it was useful to download the videos for target vowels in the current study.

Figure 1: A screenshot of one of the downloadable videos from Seeing Speech interactive website (Lawson et al., 2018)

For the training materials, seven video clips that show the tongue and the lips movement (see Figure 1), were downloaded to be used for training. Each video clip was downloaded on a separate PowerPoint slide with some example words that have the same vowel. For example, for the vowel /ɒ/, the downloadable video for the London dialect included the word *lot*; two other words that have a similar vowel were also written on the slide and practiced e.g., *box, cot.*

In each session, the instructor started the session with explaining the training procedure, starting from how opening and closing the jaw, moving the tongue backwards and forwards and lip rounding/spreading affect vowel production. Then, the participants were trained for each vowel by explaining what they can see on the video, the tongue surface that has a green

light and the ultrasound video that shows the tongue movement. The video also shows the lips movement. After each vowel was explained, the instructor asked the participant to produce the word in the video and gave feedback to adjust their production. Then after explaining a pair of vowels, the instructor compared the vowels, by asking the participants to place their hand under their jaw to notice a jaw dropping between two vowels (e.g., /e/-/ɪ/) or by comparing the jaw movement and the lip rounding as in (e.g., /ɔː/-/ɑː/). This procedure was followed till all the target vowels were covered. Then the session was completed by a revision of all the vowels. The same procedure was followed in all 3 sessions, with each session lasting 45 minutes.

3. Results

3.1 Vowel Identification

Figure 2 shows the accuracy in vowel identification task by participants in the training, and the control groups, and it seems that the training group changed their performance after training. To assess any significant change after training, a linear mixed model effect was fit with the accuracy, test, and groups as fixed factors with random slope of tests (pre & post). The model showed that there was a significant effect of group, $\chi 2$ (1) =9.11, $p<.01$, indicating a difference between the training and the control group. The orthogonal contrasts showed that the training group performed better than the control group, b =-0.038, SE =0.0167, z =-2.298, $p<.05$.

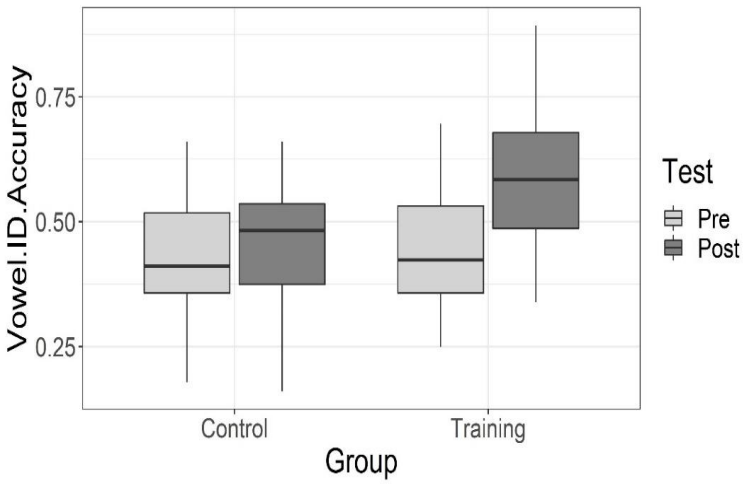

Figure 2: Proportion correct for the vowel identification at the pre- and post-tests for the training and the control group

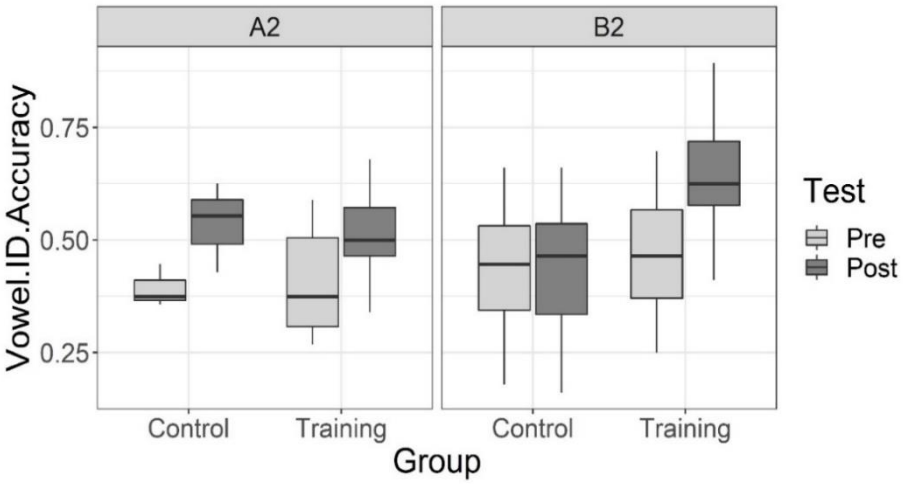

Figure 3: Boxplots showing the proportion correct for the vowel identification at the pre- and post-tests, for both proficiency levels B2 (higher level), A2 (lower level) in the training and the control group.

There was also a significant effect of test, $\chi 2$ (1) =35.4, p< .01, which shows that he participants' performance has changed after training. The orthogonal contrasts showed that participants' accuracy level was higher at the post test, b =0.142, SE =0.067, z =2.11, p<.05. Interestingly, there was a significant effect of the interaction between the group and the test, $\chi 2$ (1) =16.58, p<.001. The contrasts showed that the participants in the training group improved at the post-test more than those in the control group, b =0.0403, SE =0.009, z =4.51, p<.001. Although there was no significant effect of proficiency, there was a three-way interaction between group, test, and proficiency $\chi 2$ (1) =6.51, p<.05, showing that the high proficient learners in B2 proficiency group have performed significantly better in the post test after training compared to the other learners, b =0.21, SE =0.085, z =2.55, p<.05. It also indicates that learners with high proficiency might benefit more from the training (see Figure 4).

3.2 Category Discrimination

Figure 4 shows the accuracy proportion for the category discrimination task before and after the training for the training and the control group. The figure shows that the participants in both groups have changed their performance with an advantage for the group which was trained.

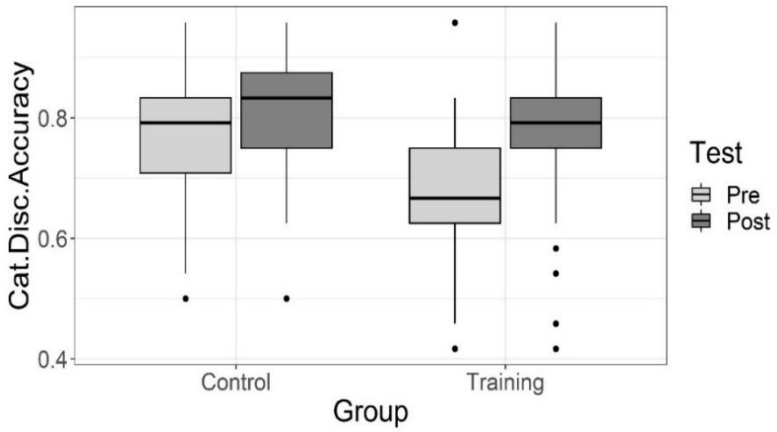

Figure 5: Boxplots of Category discrimination accuracy at the pre- and post-tests for the participants in both groups, the training, and the control group.

In order to test this observation, a linear mixed model was built for the category discrimination data. The best-fitting model included test (pre & post), and group (training & control) as fixed factors with a random intercept of participants. The model showed that there was a significant effect of test, $\chi 2$ (1) =11.69, $p<.001$, indicating a difference between the pre- and post-tests. The model also showed a significant effect of group, $\chi 2$ (1) =4.045, $p<.05$. The factor contrasts showed that the training group performed better than the control group, b =-0.081, SE =0.0338, z =-2.410, $p<.05$. However, there was no significant interaction between test and group, $p>.05$.

3.3 Vowel Intelligibility

Figure 5 shows that participants in the training group were more intelligible after training compared to those in the control group. In order to assess any significant improvement after training, a linear mixed effects model was built with the intelligibility data.

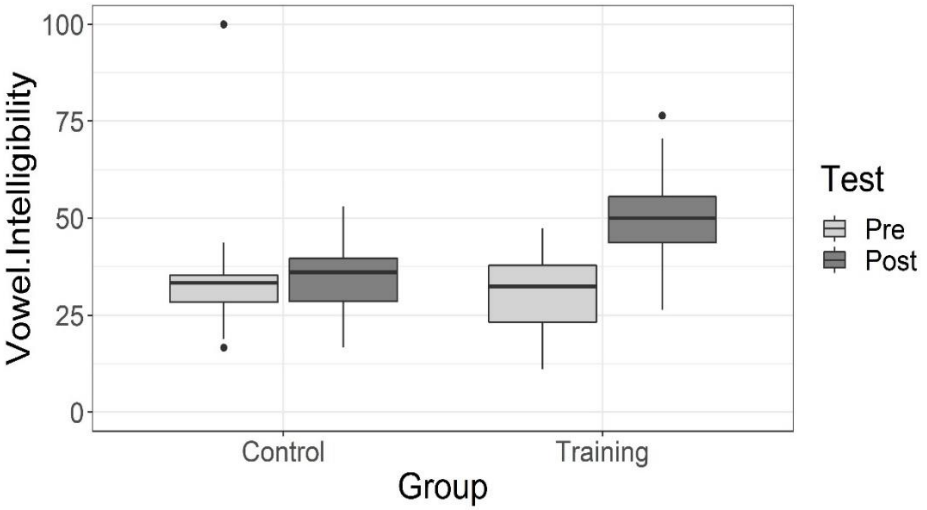

Figure 5: Boxplots of the vowel intelligibility (how accurate non-native speakers' word production, was identified by native SSBE speakers) at the pre- and post-test for the control

and the training group.

The best-fitting model included group (control & training) and test (pre & post) as fixed factors with random intercept of the speaker (non-native speaker). The model showed a significant effect of test $\chi 2$ (1) =15.707, p<.01. The effect of the group was significant $\chi 2$ (1) =15.98, p<.01, and there was a significant interaction between group and test, $\chi 2$ (1) =8.904, p<.01. The contrast shows that participants in the training group performed better in the post test than the control group, b =1.054, SE =3.53, z =2.98, p<.01. This indicates that only after 3 sessions of training, participants production improved to be more intelligible.

3.4 The Correlation Between Improvement in Perception and Production

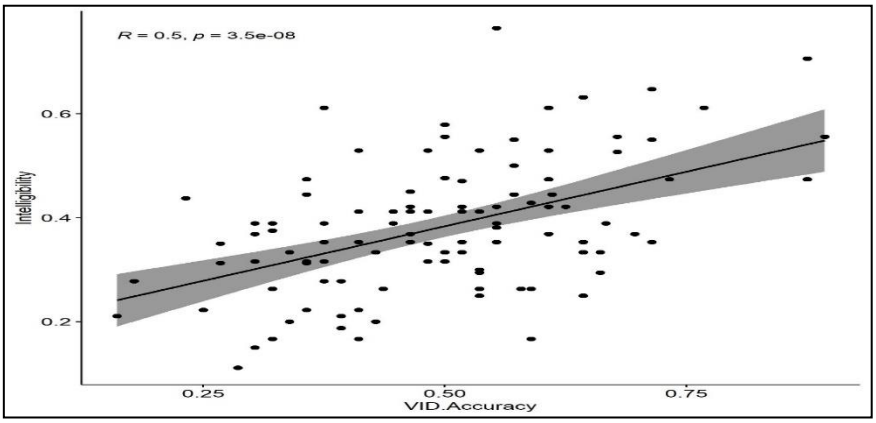

Figure 6: Scatterplot showing the correlation between vowel intelligibility and vowel identification overall (the average between the intelligibility at the pre- and the post-test) and vowel identification scores (the average of the vowel identification tasks at the pre- and the post-test).

Figure 6 displays the correlation between the vowel identification accuracy and the intelligibility scores. A Pearson's correlation indicated a significant positive correlation between the vowel identification accuracy and that of the intelligibility, [r=0.496, p<.001, R^2 =0.24], indicating that the performance in production is reflected on the learners' perception accuracy. That is, the participants who produced the vowels accurately, scored higher accuracy level for vowel identification, while those who were less intelligible, scored lower vowel identification accuracy. This correlation was found to be stronger in the post test, [r=0.515, p<.001, R^2 =0.26], than in the pre-test, [r=0.215, p>.05, R^2 =0.04] (see Figure 7).

Moreover, when comparing the correlation between the vowel identification and the intelligibility for each group, we found no significant corelation for the control group, [r=0.18, p>.05, R^2 =0.03]. However, for the training group, the correlation between the two variables was significant, [r=0.544, p<.001, R^2 =0.29]. These correlations indicate that training helped participants to form some sort of relationship between the vowel perception and the production imbedded in words.

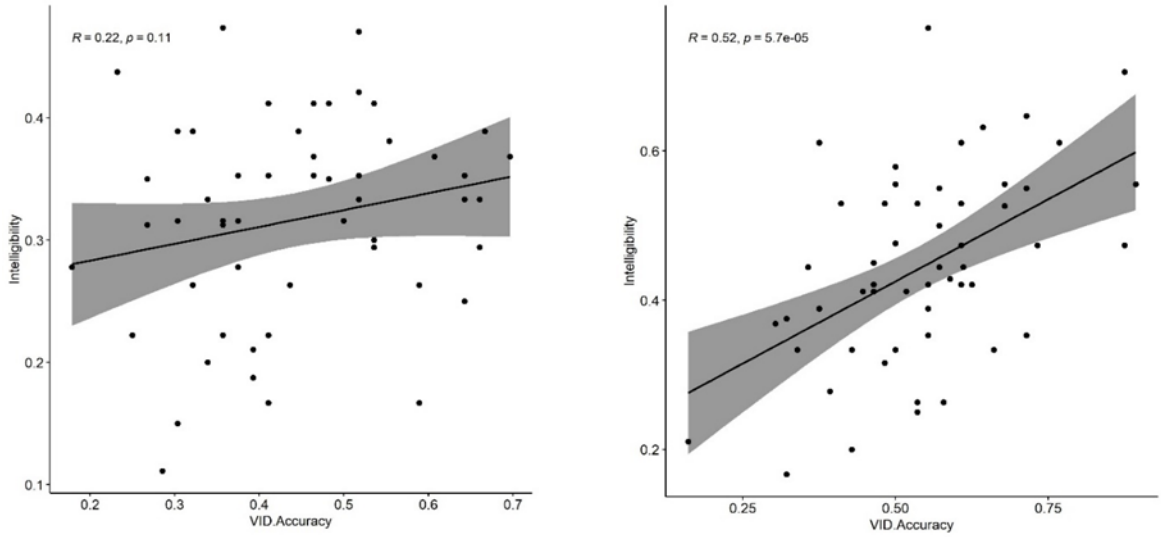

Figure 7: Scatterplot showing the correlation between vowel intelligibility and vowel identification overall at the pre-test (on the left-hand side) and at the post-test (on the right-hand side)

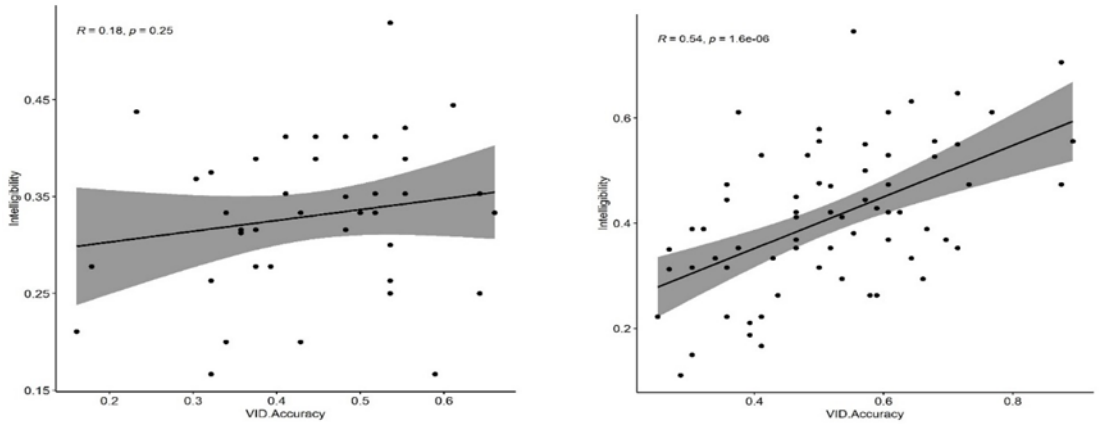

Figure 8: Scatterplot showing the correlation between vowel intelligibility and vowel identification overall for the control group (on the left-hand side) and for the training group (on the right-hand side)

4. Discussion

4.1 The Effect of Training on Vowel Production

This study examined the potential benefits of using online articulatory resources along with audio instructions and feedback in improving L2 vowel production, and whether learning proceeds to improving vowel perception. Native Arabic speakers were trained over three sessions to produce seven SSBE vowels embedded in single words. The UTI with lip videos were used as a tool to show participants how to produce the target vowels by moving their tongue and lips similar to that displayed on the screen while receiving audio instruction and feedback from the instructor. Production was assessed by intelligibility of word production

before and after the training. The results of the production showed that after three sessions of training, learners in the training group were found to be more intelligible than their production at the pre-test, indicating that their vowel production has improved as a result of the training compared to the control group. These results are in line with previous studies which found that L2 adult learners show some benefits of articulatory training in producing L2 speech sounds (e.g., Sisinni et al., 2016; Massaro et al., 2008; Carey, 2004; Kartushina et al., 2015). The improvement in the production indicates that the tool used in the training is beneficial for vowel learning, especially as they could see the front, back and the height of the tongue from the palate which was the main reason for choosing this UTI source. This visualisation was accompanied by instructions from the instructor who emphasised a jaw drop, asking them to place their hand under their jaw when producing the vowel contrast to notice the difference, when the tongue height changes between /ʊ-ɒ/, /ɪ-e/ or emphasizing the difference in lip rounding or spreading /ʊ-ʌ/, /ɑː-ɔː/. For the /ɪ-e / contrast, learners might have found it difficult at first as they do not have this contrast in their L1 and therefore, assimilated both vowels in to one Arabic vowel /i/ as proposed by PAM (Best, 1995). This in turn might explain their difficulty with forming new L2 categories as these vowels are close to L1 vowels (Flege, 1995). The same can be said about the /ʊ-ɒ/ vowels as they can be assimilated into the Arabic back vowel /u/. Hence, before training, the participants of the current study did not seem to be aware of how differently the English vowels are produced. However, after training and associating tongue movement to these vowels, the participants' attention might have shifted to detect the differences between vowel pairs, and thus they were found to be more intelligible after training. This shows the importance of including articulatory instructions in teaching. Even if L2 learners are exposed to a certain English variant, some articulatory instructions to clarify the difference between some categories might improve their pronunciation. Such instructions may enable learners to correctly produce phonemes that make them more intelligible and perceive speech accurately, so they can get the most of conversations in or outside a classroom setting.

4.2 The Effect of Production Training on Vowel Perception

The results also showed that Learners' perception has improved after training. In contrast to the notion that suggests that production of L2 phonemes disrupts perceptual learning (cf., Baese-Berk & Samuel, 2016), participants who received production training showed improvement in both perceptual tasks, vowel identification, and the vowel category discrimination, after training. Similar results have been found in studies where articulatory training improves L2 phoneme perception (e.g., Kartushina et al., 2015; Cibelli, 2022), suggesting that production training may tune the corresponding perceptual representations of speech sounds. In fact, some studies (e.g., Lametti et al., 2014) found that the motor process of speech production leads to perceptual changes that are associated with speech motor learning. The result of transfer across speech modalities is in line with the Direct Realism Theory (Best, 1995) which states that listeners perceive speech as an articulatory gesture (e.g., tongue, lip, and jaw movement) of speakers. In a similar way, the Motor theory of speech perception suggests that speech is perceived with speech-specific cognitive module (phonetic module) that detects the intended articulatory gesture of the speaker (Liberman & Mattingly, 1985). These theories propose a link between speech perception and production which is compatible with

the positive correlation between vowel perception and vowel production, which was measured by the intelligibility task, suggesting that the individuals who were found to be intelligible after training, were also the same individuals whose perception has improved. The correlation had the same pattern when run only with the training group, R =.05, p<.05. However, the correlation was not significant for the control group (see Figure 8). This indicates that training played a role in learning, as the learners seemed to make a connection between the production and the perception of the vowels. This correlation between the two speech modalities was found to be weak in different studies (e.g., Bradlow et al., 1997; Hattori, 2010; Peperkamp & Bouchon, 2011; Wong, 2013; Alshangiti& Evans, 2014) where training on one modality did not proceed to improvement of the other or showed little change. For example, Alshangiti & Evans (2014) found that training is domain specific, i.e., when learners are trained on production, their production may improve, but not their perception. Although their study trained Arabic speakers on vowel production, the amount of improvement in vowel learning is more noticeable in our study, which is possibly due to the number of vowels trained (14) compared to vowels in our study (7). They argued following Nishi and Kewley-Port (2007) that training on vowels is better with a whole set of vowels rather than a subset. While in the current study we did not measure generalisation to untrained vowels, the smaller set of vowels in our study seemed to help learners focus on limited contrasts and that might have helped them improve their performance in producing and perceiving the trained vowels after training. Another possibility of different training output is probably the tool of training used as learners in Alshangiti & Evans (2014) study were trained face to face with a tutor with the help of computer assisted vowel training interface. In our study we had audio instructions to avoid excessive visual information, so the participants can focus on the information given on the UTI videos while listening to instructions and feedback. It is possible then that the audio instructions had extra perceptual input that helped learners improve their perception. Another difference that might explain the difference between training output in the current study with that of Alshangiti and Evans (2014) is that their participants were residents of an English-speaking country during training, while the participants in this study live in a non-immersion setting, and thus might have relied on the training input, at least during the training days, to improve their production. This suggests that training L2 production with reliable online resources while receiving individual feedback and getting instructions that suit individual learners can be useful specifically for L2 learners in non-immersion settings.

4.3 The Effect of Learners' Proficiency Level on Performance

Learners' level of proficiency seemed to have some effect on their performance. Although all the learners improved their perception after training, the high-level learners showed better results after the training. Initially, before the training participants in both the training and control group had similar mean accuracy. However, Figure 3 shows that the control group had some improvement, although not significant, the low proficient (A2) group seemed to learn something about the stimuli merely by repeated exposure to the stimuli. The same exposure effect seemed to help the high proficiency learners in the control group, as mentioned earlier, most participants in the control group are from higher proficiency level (B2) and therefore their starting point at the pre-test, specifically in the category discrimination, was above chance and

by exposure to the stimuli repetition, they got some improvement, though not significant. On the other hand, the training group found to improve significantly in both perceptual tasks, indicating that not only exposure to the stimuli that played a role in their improvement, but also the training method with UTI along with the audio instructions and feedback which found to be crucial in improving L2 pronunciation (cf. Saito & Lyster, 2012). In sum, our results suggest that pronunciation training is beneficial for learners regardless of their proficiency level. Nevertheless, the high proficient learners may have used more clues as they already have the lexical knowledge to help them make use of differences between some words that they have not noticed, before training, that they confuse their pronunciation.

4.4 Limitation and Further Research

The training method that was used in this study by were some online resources which were effective at improving vowel production and showed some evidence of improvement in vowel perception. However, given that we only trained a subset of SSBE vowels (cf. Nishi and Kewley-Port, 2007), it is not clear from our results whether learning can be generalised to untrained vowels, and whether the learning can be retained after training. Future work might compare production training on whole versus sub-set of vowels and test learning generalisation to untrained vowels.

Production accuracy in our study was evaluated by native speakers' identification of L2 vowel production, which might be biased by raters and their experience in perceiving L2 speech. That being said, other objective measures like Mahalanobis distance (Kartushina et al., 2015) or acoustic measurements (Alshangiti & Evans, 2014) proved to be valuable for the assessment of production performance (cf. Delvaux et al., 2013). Ideally, combining both objective and subjective measures may provide a complementary approach to assess L2 production accuracy.

Although participants improved after training, it was below expectations which might have to do with the insufficient training sessions. It is possible that with more training sessions, learners' performance would increase. Moreover, in training sessions, it is ideal if there are continuous tests after each session to track learners' progress to see when their learning plateaued out to determine how many sessions a certain level of proficiency might need for learning.

5. Conclusion

This study has shown that online resources for pronunciation training can be beneficial for improving L2 production if they are accompanied by adequate instructions and feedback. Moreover, the results showed some transfer of production learning to perception, which was approved by a positive correlation between vowel perception and production only in the training group. Even though participants in the control group were from a higher-level proficiency, the training group showed evidence of learning that linked the two speech modalities. This link is compatible with some speech theories that suggest that speech perception is intended or produced gestures (Best, 1995; Liberman & Mattingly, 1985). Our training approach with visual aids and online audio feedback can help students who feel shy in classrooms and give teachers the option of teaching some skills online where they can use some

assisted learning tools. This method can help L2 learners to scaffold their pronunciation with ease and freedom to ask about individual difficulties that might not be possible during class time. In our study, we found some benefits of using UTI videos in training, though more work is needed to further validate the efficacy of online resources that provide articulatory visual aids, and whether audio instructions make a difference.

Acknowledgments

Acknowledgments to the participants for their cooperation.

References

Abel, J., Allen, B., Burton, S., Kazama, M., Kim, B., Noguchi, M., ... & Gick, B. (2015). Ultrasound-enhanced multimodal approaches to pronunciation teaching and learning. *Canadian Acoustics*, *43*(3), 124-125.

Akahane-Yamada, R., McDermott, E., Adachi, T., Kawahara, H., & Pruitt, J. S. (1998). Computer-based second language production training by using spectrographic representation and HMM-based speech recognition scores. In *Fifth International Conference on Spoken Language Processing*.

Al-Ani, S. H. (2014). Arabic phonology. In *Arabic Phonology*. De Gruyter Mouton. https://doi.org/10.1515/9783110878769

Alshangiti, W., & Evans, B. G. (2014, May). Investigating the domain-specificity of phonetic training for second language learning: comparing the effects of production and perception training on the acquisition of English vowels by Arabic learners of English. In *The Proceedings of the International Seminar for Speech Production*.

Alshangiti, W., Evans, B., & Wibrow, M. (2019). Learning to speak in a second language: Does multiple talker production training benefit production of English vowels in Arabic children?. In *Proceedings of the International Congress of Phonetic Sciences* (pp. 2193-2197). International Phonetic Association.

Baese-Berk, M. M., & Samuel, A. G. (2016). Listeners beware: Speech production may be bad for learning speech sounds. *Journal of Memory and Language*, *89*, 23-36. https://doi.org/10.1016/j.jml.2015.10.008

Bälter, O., Engwall, O., Öster, A. M., & Kjellström, H. (2005, October). Wizard-of-Oz test of ARTUR: a computer-based speech training system with articulation correction. In *Proceedings of the 7th international ACM SIGACCESS conference on Computers and accessibility* (pp. 36-43).

Bangun, I., & Liontas, J. I. (2019). Sounds of Speech Published: iOS, Android-Jerry Moon, University of Iowa Institute Service, Information Technology Service University of Iowa, Carnegie. *The Reading Matrix: An International Online Journal*, *19*(1).

Best, C. T. (1995). A direct realist view of cross-language speech perception. *Speech Perception and Linguistic Experience*, 171-206.

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: commonalities and complementarities. Lang Exp. *Second Lang. Speech Learn*, *1334*, 1-47.

Bliss, H., Abel, J., & Gick, B. (2018). Computer-assisted visual articulation feedback in L2 pronunciation instruction: A review. *Journal of Second Language Pronunciation*, *4*(1), 129-153. https://doi.org/10.1075/jslp.00006.bli

Boersma, P.& Weenink, D. (2022). Praat: Doing phonetics by computer [Computer program]. Version 6.2.14, retrieved 24 May 2022 from http://www.praat.org

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*(4), 2299-2310. https://doi.org/10.1121/1.418276

Campbell, F. M. (2004). *The gestural organization of North American English/r: A study of timing and magnitude* (Doctoral dissertation, University of British Columbia).

Camus, P. (2019). The effects of explicit pronunciation instruction on the production of second language Spanish voiceless stops: a classroom study. *Instructed Second Language Acquisition*, *3*(1), 81-103. https://doi.org/10.1558/isla.37279

Carey, M. (2004). CALL visual feedback for pronunciation of vowels: Kay Sona-Match. *CALICO Journal*, 571-601.

Carley, P., Mees, I. M., & Collins, B. (2017). *English phonetics and pronunciation practice*. Routledge.

Cebrian, J. (2007, January). Old sounds in new contrasts: L2 production of the English tense-lax vowel distinction. In *Proceedings of the 16th International Congress of Phonetic Sciences* (Vol. 3, pp. 1637-1640).

Cibelli, E. (2022). Articulatory and perceptual cues to non-native phoneme perception: Cross-modal training for early learners. *Second Language Research*, *38*(1), 117-147. https://doi.org/10.1177/0267658320921217

d'Apolito, S., Sisinni, B., Grimaldi, M., & Fivela, B. G. (2017). Perceptual and ultrasound articulatory training effects on English L2 vowels production by Italian learners. *World Academy of Science, Engineering and Technology International Journal of Cognitive and Language Science*, *11*, 2447-2453.

Delvaux, V., Huet, K., Piccaluga, M., & Harmegnies, B. (2013). Production training in second language acquisition: a comparison between objective measures and subjective judgments. In *INTERSPEECH* (pp. 2375-2379).

Derwing, T. M., & Munro, M. J. (2015). *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research* (Vol. 42). John Benjamins Publishing Company.

Dowd, A., Smith, J., & Wolfe, J. (1998). Learning to pronounce vowel sounds in a foreign language using acoustic measurements of the vocal tract as feedback in real time. *Language and Speech*, *41*(1), 1-20. https://doi.org/10.1177/002383099804100101

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

Engwall, O., & Bälter, O. (2007). Pronunciation feedback from real and virtual language teachers. *Computer Assisted Language Learning*, *20*(3), 235-262. https://doi.org/10.1080/09588220701489507

Escudero, P., & Boersma, P. (2004). Bridging the gap between L2 speech perception research and phonological theory. *Studies in Second Language Acquisition*, *26*(4), 551-585. https://doi.org/10.1017/S0272263104040021

Evans, B. G., & Alshangiti, W. (2018). The perception and production of British English vowels and consonants by Arabic learners of English. *Journal of Phonetics*, *68*, 15-31. https://doi.org/10.1016/j.wocn.2018.01.002

Feng, Z. (2020). Effects of Identification and Pronunciation Training Methods on L2 Speech Perception and Production: Training Adult Japanese Speakers to Perceive and Produce English/r/-/l/. *Studies in Applied Linguistics & TESOL*, *20*(2), 57-83.

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. *Speech perception and linguistic experience: Issues in cross-language research*, *92*, 233-277.

Flege, J. E., & Bohn, O. S. (2021). The revised speech learning model (SLM-r). *Second language speech learning: Theoretical and empirical progress*, 3-83.

Flege, J. E., MacKay, I. R., & Meador, D. (1999). Native Italian speakers' perception and production of English vowels. *The Journal of the Acoustical Society of America*, *106*(5), 2973-2987. https://doi.org/10.1121/1.428116

Gick, B., Bernhardt, B., Bacsfalvi, P., Wilson, I., & Zampini, M. (2008). Ultrasound imaging applications in second language acquisition. *Phonology and second language acquisition*, *36*(6), 309-322.

Gilakjani, A. P. (2012). A study of factors affecting EFL learners' English pronunciation learning and the strategies for instruction. *International Journal of Humanities and Social Science*, *2*(3), 119-128.

Hattori, K. (2010). *Perception and production of English/r/-/l/by adult Japanese speakers* (Doctoral dissertation, UCL (University College London)).

Kaiser, D. J. (2017). iPronounce: Understanding pronunciation apps. *Online Webinar*, *8*.

Kartushina, N., & Martin, C. D. (2019). Talker and acoustic variability in learning to produce nonnative sounds: Evidence from articulatory training. *Language Learning*, *69*(1), 71-105. https://doi.org/10.1111/lang.12315

Kartushina, N., Hervais-Adelman, A., Frauenfelder, U. H., & Golestani, N. (2015). The effect of phonetic production training with visual feedback on the perception and production of foreign speech sounds. *The Journal of the Acoustical Society of America*, *138*(2), 817-832. https://doi.org/10.1121/1.4926561

Kartushina, N., Hervais-Adelman, A., Frauenfelder, U. H., & Golestani, N. (2016). Mutual influences between native and non-native vowels in production: Evidence from short-term visual articulatory feedback training. *Journal of Phonetics*, *57*, 21-39. https://doi.org/10.1016/j.wocn.2016.05.001

Lametti, D. R., Rochet-Capellan, A., Neufeld, E., Shiller, D. M., & Ostry, D. J. (2014). Plasticity in the human speech motor system drives changes in speech perception. *Journal of Neuroscience*, *34*(31), 10339-10346. https://doi.org/10.1523/JNEUROSCI.0108-14.2014

Lawson, E., Stuart-Smith, J., Scobbie, J. M., Nakai, S., Beavan, D., Edmonds, F., ... & Durham, M. (2015). Dynamic Dialects: an articulatory web resource for the study of accents [website].

Lengeris, A. (2009). Perceptual assimilation and L2 learning: Evidence from the perception of Southern British English vowels by native speakers of Greek and Japanese. *Phonetica*, *66*(3), 169-187. https://doi.org/10.1159/000235659

Lengeris, A. (2018). Computer-based auditory training improves second-language vowel production in spontaneous speech. *The Journal of the Acoustical Society of America*, *144*(3), EL165-EL171. https://doi.org/10.1121/1.5052201

Levis, John M. "Changing contexts and shifting paradigms in pronunciation teaching." *TESOL Quarterly* 39, no. 3 (2005): 369-377. https://doi.org/10.2307/3588485

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

Massaro, D. W., Bigler, S., Chen, T., Perlman, M., & Ouni, S. (2008). Pronunciation training: the role of eye and ear. In *Ninth Annual Conference of the International Speech Communication Association*.

McCrocklin, S. M. (2016). Pronunciation learner autonomy: The potential of automatic speech recognition. *System*, *57*, 25-42. https://doi.org/10.1016/j.system.2015.12.013

Mozaffari, M. H., Wen, S., Wang, N., & Lee, W. S. (2019). Real-time Automatic Tongue Contour Tracking in Ultrasound Video for Guided Pronunciation Training. In *VISIGRAPP (1: GRAPP)* (pp. 302-309).

Neri, A., Cucchiarini, C., & Strik, H. (2008). The effectiveness of computer-based speech corrective feedback for improving segmental quality in L2 Dutch. *ReCALL*, *20*(2), 225-243. https://doi.org/10.1017/S0958344008000724

Nishi, K. & Kewley-Port, D. (2007). Training Japanese Listeners to Perceive American English Vowels: Influence of Training Sets *Journal of Speech, Language, and Hearing Research* 50(6): 1496–1509. https://doi.org/10.1044/1092-4388(2007/103)

Olson, D. J. (2014). Phonetics and technology in the classroom: A practical approach to using speech analysis software in second-language pronunciation instruction. *Hispania*, 47-68.

Peperkamp, S., & Bouchon, C. (2011). The relation between perception and production in L2 phonological processing. In *Twelfth Annual Conference of the International Speech Communication Association*.

Pillot-Loiseau, C., Antolík, T. K., & Kamiyama, T. (2013, August). Contribution of ultrasound visualisation to improving the production of the French/y/-/u/contrast by four Japanese learners. In *PPLC13: Phonetics, Phonology, Languages in Contact Contact: Varieties, Multilingualism, Second Language Learning*.

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

Shinohara, Y., & Iverson, P. (2018). High variability identification and discrimination training for Japanese speakers learning English /r/–/l. *Journal of Phonetics*, *66*, 242-251. https://doi.org/10.1016/j.wocn.2017.11.002

Sisinni, B., d'Apolito, S., Fivela, B. G., & Grimaldi, M. (2016). Ultrasound articulatory training for teaching pronunciation of L2 vowels. In *Conference Proceedings. ICT for Language Learning* (pp. 265-270). libreriauniversitaria. it Edizioni.

Thomson, R. I. (2011). Computer assisted pronunciation training: Targeting second language vowel perception improves pronunciation. *CALICO Journal*, *28*(3), 744-765.

Thomson, R. I., & Derwing, T. M. (2015). The effectiveness of L2 pronunciation instruction: A narrative review. *Applied Linguistics*, *36*(3), 326-344. https://doi.org/10.1093/applin/amu076

Watson, J. C. (2007). *The phonology and morphology of Arabic*. OUP Oxford.

Wik, P. (2011). *The virtual language teacher* (Doctoral dissertation, Ph. D. thesis, KTH School of Computer Science and Communication).

Wilson, I., Gick, B., O'Brien, M. G., Shea, C., & Archibald, J. (2006). Ultrasound technology and second language acquisition research. In *Proceedings of the 8th Generative Approaches to Second Language Acquisition Conference (GASLA 2006)* (pp. 148-152). Somerville, MA: Cascadilla Proceedings Project.

Wong, J. W. S. (2013, August). The effects of perceptual and/or productive training on the perception and production of English vowels/ɪ/and/iː/by Cantonese ESL learners. In *Interspeech* (pp. 2113-2117).

Yoshida, M. T. (2018). Choosing technology tools to meet pronunciation teaching and learning goals. *CATESOL Journal*, *30*(1), 195-212.