Learning English vowels: the effects of different phonetic training modes on Arabic learners' production and perception. [1]

Wafaa Alshangiti[1] and Bronwen G. Evans [2]

[1] *English Language Institute, King Abdulaziz University, Jeddah, 21589, Saudi Arabia.*

walshangiti@kau.edu.sa

[2] *Speech, Hearing and Phonetic Sciences, University College London, London, WC1N 1PF, United Kingdom.*

bronwen.evans@ucl.ac.uk

This study investigated the effect of different types of phonetic training on potential changes in the production and perception of English vowels by Arabic learners of English. Forty-six Arabic learners of English were randomly assigned to one of three high variability vowel training programs: Perception training (High Variability Phonetic Training: HVPT), Production training, and a Hybrid Training program (production and perception training). Pre- and post-tests (vowel identification, category discrimination, speech recognition in noise and vowel production) showed that all training types led to improvements in perception and production. There was some evidence that improvements were linked to training type: learners in the Perception Training condition improved in vowel identification but not vowel production, whilst those in the Production Training condition showed only small improvements in performance on perceptual tasks, but greater improvement in production. However, effects of training modality were complicated by proficiency, with high proficiency learners benefitting more from different types of training regardless of training mode than lower proficiency learners.

[1] Portions of this work were presented at the International Seminar for Speech Production. Cologne, Germany, May 5th-8th 2014.

I.INTRODUCTION

Learning to accurately perceive and produce the sounds of a second language (L2) in adulthood is well-known to be challenging. Although some learners go on to achieve a high degree of fluency, many find learning an L2 difficult and often speak their L2 with a noticeable foreign accent. For example, Saudi Arabic learners, the focus of this study, find acquiring English vowels particularly difficult (Evans & Alshangiti, 2018; Alshangiti, 2015). Arabic languages typically have a small vowel inventory: although the exact number varies according to the dialect, Modern Standard Arabic, the standard variety used across the Arabic-speaking world and the language of education, is described as having 3 tense-lax monophthong pairs, /i/-/iː/, /a/-/aː/, /u/-/uː/ (Holes, 2004). Hijazi Arabic, the variety spoken in Saudi Arabia, has two additional monophthongs, /eː/ and /oː/ and two diphthongs, /aj/ and /aw/ (Watson, 2007; Jarrah, 1993). In contrast, Standard Southern British English (SSBE) is typically described as having 12 monophthongs and 8 diphthongs (e.g., Wells, 1982). In a previous study (Evans & Alshangiti, 2018), we showed that although some English vowels are perceived accurately by beginners, e.g., *heed* /iː/, *head* /e/, Arabic learners of English find others much harder to identify even when they have a high degree of proficiency in English. For example, *hid* /ɪ/ was frequently confused with *head* /e/ and the central and low back vowels /ɒ ʌ əʊ/ were poorly categorized by both high and low proficiency learners (cf. Shafiro et al., 2012). Likewise, these vowels were produced least accurately: all participants were also least accurate in their production of the high front vowel contrast *hid-head* /ɪ-e/, the low central vowels *hod* and *hud* /ɒ-ʌ/, and the diphthong *hoed* /əʊ/. This is consistent with the hypothesis that L2 production and perception co-evolve (e.g., Flege & Bohn, 2021). In the current study, we investigate the effect of three different training paradigms – Perception Training, Production Training and a Hybrid Training (production & perception) program – on the perception and production of English vowels by Arabic learners of English.

A: Background

Previous work has suggested that difficulties in L2 perception and production arise as a result of the relationship between the L1 and L2 (e.g., SLM-r, Flege & Bohn, 2021; PAM-L2, Best & Tyler, 2007). That is, L2 learners find it difficult to acquire new phonetic categories not because there is a change in the mechanisms through which new sounds are learned (Flege & Bohn, 2021), but because experience with the L1 alters low-level processing, and these changes interfere with their ability to alter existing representations and to form new categories for L2 sounds (Iverson et al., 2003). L2 speech learning is thus thought to be shaped by perceptual biases induced by the L1 phonetic system (Flege & Bohn, 2021; Best & Tyler, 2007).

In the SLM-r, Flege & Bohn (2021) suggest that L2 learners make use of the same mechanisms and processes when learning their L2 as when learning their L1 as children, but that applying these mechanisms and processes to an L2 does not lead to the same outcomes. These are thought to arise for a number of reasons. First, L2 sounds are initially perceived as L1 sounds because they are linked to sounds in the L1 phonetic inventory. This means that pre-existing L1 categories interfere with and sometimes block the formation of new, L2 phonetic categories. The similarity between L2 phonemes also affects learning from the first day of leaning L2 (Flege, 1995). This idea is further developed in the PAM-L2, in which Best & Tyler (2007) suggest that identification and discrimination of L2 phonemes is not just determined by L1-L2 relationships, but by how contrasting L2 phonemes relate to each other within the emerging L1-L2 phonological space. Second, the SLM-r argues that L2 learning differs from L1 learning because the input that is received by monolingual native speakers of the language may differ from that received by L2 learners when learning the same sounds. Input may differ in terms of the amount: L2 learners, even those living in the L2-speaking country, typically continue to use their L1 so will likely not receive as much input as monolingual children who need to reach adult-like performance levels. The input may

also differ in quality: for example, learners may be more likely to hear foreign-accented versions of their L2 because they are taught by non-native speakers, or because they are more likely to use their L2 in multilingual environments where they are exposed to their native and non-native accented varieties of their L2.

Despite this, a lot of evidence from studies of L2 learning shows that increased exposure to L2 phonetic categories can improve perception and production of difficult non-native contrasts (see e.g., Flege & Bohn, 2021 for a review). This input can come from real-world experience (e.g., Flege et al.,1995) or from targeted invention, in particular phonetic training (e.g., Logan et al., 1991). The focus of phonetic training paradigms has been largely on training perception with High Variability Phonetic Training (HVPT), in which learners are exposed to multiple instances of L2 phonetic categories produced by multiple talkers in a variety of phonetic contexts, particularly successful in improving phoneme identification (Logan et al., 1991; Bradlow et al., 1997; Iverson & Evans, 2009; Shinohara & Iverson, 2018). Much of this work has focused on training consonants (e.g., Logan et al., 1991; Bradlow et al., 1997), though more recently, a number of studies have adapted the HVPT method for training vowels in English (see e.g., Lambacher et al., 2005 for Japanese learners; Nishi & Kewley-Port, 2007 for Japanese learners; Iverson & Evans, 2009 for Spanish & German learners; Thomson, 2012 for Mandarin learners) and other languages (e.g., Inceoglu, 2016 for English learners of French). These studies typically show an increase in performance of 16-25 percentage points on vowel intelligibility tasks, with training on larger sets of vowels more effective than that concentrating on a limited number of the most difficult vowels (Nishi & Kewley-Port, 2007; cf. Sakai & Moorman, 2018). Although it is clear that there is an HVPT benefit for perception, it is unclear whether this transfers to production. In a meta-analysis of 30 phonetic training studies, Sakai & Moorman (2018) find that learners who are trained in perception experience medium-sized improvements in perception but that even though there are robust gains in production, these are

only small and are highly variable across studies. Indeed, some studies show that perceptual training using HVPT leads to improvements in both speech perception and production (e.g., Bradlow et al., 1997; Lengeris & Hazan, 2010; Shinohara & Iverson, 2018), but others find little or no transfer, with training in perception yielding no improvement in production (e.g., Hattori & Iverson, 2009; Hanulíková et al., 2012).

Results from studies that have trained production are likewise mixed. As for perception, some have found that production training leads to large gains in production accuracy but does not transfer to perception, whilst others have found that production training improves both production and perception. For example, Hattori (2010) found that after ten sessions of one-to-one pronunciation training in which they listened to and saw spectrograms of their own and native productions, Japanese speakers were able to produce native-like /ɹ/-/l/, but that their perception of this contrast did not improve. In contrast, Kartushina et al. (2015) trained L1 French speakers on two Danish vowel contrasts, /e/-/ɛ/ and /y/-/ø/, that are not present in French. Participants completed five, 45mins training sessions, with one hour of training per vowel. Feedback was based on real-time, trial-by-trial acoustic analysis of participants' vowel productions, which were compared to that of native Danish speakers. Feedback was visual: participants saw a plot of their vowel and that of the native Danish speaker in an F1-F2 space. Participants improved in production for all four trained vowels and this transferred to perception. Interestingly, a control group who repeated vowels but who did not receive visual feedback showed no improvement in production or perception, leading the authors to suggest that improvements in production were driven by the visual feedback itself. However, these were naïve participants with no previous experience of Danish, with training and testing limited to isolated vowels; these improvements may thus represent initial gains made when first exposed to new phonetic categories through intensive training.

Still other studies have used a combination of perception and production training. For example, Herd et al. (2013) trained monolingual American English learners of Spanish with Spanish intervocalic /d ɾ r/. Learners were assigned to either a perception, production or combination (production & perception) high variability training programme. All 3 training paradigms were successful – all learners improved in their performance overall – but different modalities and different contrasts improved differently for the different groups. Specifically, perception trainees performed better on identification of the /d/-/ɾ/ contrast whilst production trainees performed better with the /ɾ/ -/r/ contrast. Both groups made similar improvements in production of /r/. Combination trainees made no gains in perception. All participants improved in production but there appeared to be an advantage for production and combination training, with participants in these conditions improving overall, but with those in the perception training condition improving only for /r/. Based on these results, the authors conclude that whilst perception and production training prove most effective for training perception, combination training which included only half the exposure to each modality, was most effective for training production.

But the nature of production training differs across studies (see Saito & Plonsky, 2019 for a review) and perhaps not surprisingly, so too do the outcomes of the training. For example, Baese-Berk & Samuel (2016; 2022) compared learning of a Basque contrast by naïve Spanish learners in two conditions: perception-only and perception & production (hybrid). In their initial experiment, participants were trained with the /ʂa/-/ʃa/ contrast using an ABX task in two conditions: perception and perception & production. They heard tokens synthesized along a continuum, identified them and received feedback on their responses. Perception-production training used the same task as the perception training but participants were asked to repeat the final token before making their identification. Only those given perception training improved in their identification of the contrast; those in the perception-production condition showed no increase in sensitivity to the contrast. The

authors argue that perceptual learning for those in the perception-production condition was disrupted by production (see also Leach & Samuel, 2007), not because of exposure to their own inaccurate production, but because of the combined effects of the complex set of skills involved in learning a new contrast, a process which draws not only on linguistic skills but also on cognitive abilities such as inhibitory control, working memory and attention (see e.g., Gaffarvand Mokari & Werner, 2018). This type of production training is very different from the more explicit approaches used in other studies (see Saito & Plonsky, 2019 for a review), where participants are instructed in both how to produce (articulatory information) a sound and what to listen for (auditory information), and where we typically find improvements in both production and perception (see also Sakai & Moorman, 2018).

B: The Current Study

It is notable that studies training production or using combination/hybrid approaches have typically focused on a limited number of contrasts or even just a single sound, with the majority training consonants (e.g., Hattori, 2010; Herd et al., 2013; Saito, 2013). Unlike vowels, consonants are discrete and learners can arguably benefit from somatosensory feedback in a way that is much more difficult for vowels where the articulators are in open approximation. Those studies that have trained vowels, often focus on naïve learners and intensively train a very limited number of contrasts (e.g., Kartushina et al., 2015; Kartushina & Martin, 2018; Lambacher et al., 2005). However, perceptual learning of vowels has been shown to be better when a larger number of contrasts are trained (Nishi & Kewley-Port, 2007).

The current study aims to extend previous work examining the efficacy of different high variability training paradigms for acquisition of L2 vowels in a group of Arabic learners of English, one of the largest but least studied groups of L2 English learners. Participants were assigned to one of 3 training conditions – Perception Training (i.e., classic HVPT), Production Training and Hybrid

Training (production & perception) training – with improvements in vowel perception and production measured using a battery of tests in a pre-/post-test design (see Table 1 for a summary of tasks). The number and duration of training sessions was the same across conditions.

Perception Training used the UCL Vowel Trainer (Iverson & Evans, 2009). Production Training combined 1-to-1 in-person instruction with basic articulatory phonetics training supported by CALVin (Computer Assisted Learning for Vowels interface), a custom-built computer program developed by the first-author. During training, participants received explicit instruction about how to produce a sound, viewed animations, and heard and imitated examples of target vowels embedded in words. This meant that our Production Training included listening to speech, but unlike in perception training, participants were not directed to listen in a particular way, either explicitly or through the design of the task (e.g., as in HVPT). We took this approach for a number of reasons. Visual feedback has typically involved spectrogram feedback (e.g., Hattori, 2010) or for vowels, an F1-F2 plot (Kartushina et al., 2015). However, we felt that learning to recognize spectrograms or having enough understanding of the articulatory-acoustic relationship to interpret an F1-F2 plot for a large number of vowels would likely be difficult for learners with no background in phonetics. We also wanted to develop and test an approach which would be usable within the L2 teaching world, i.e., by teachers with no training in acoustic phonetics where pronunciation training typically involves imitating and therefore listening to your own and a teacher's production (see e.g., Trofimovich & Gatbonton, 2006). Other studies have used automatic speech recognition (ASR) based computer-assisted language learning (CALL) systems (e.g., Neri et al., 2008), combining these with virtual talker heads, e.g., Baldi (Massaro & Light, 2003) or a virtual language teacher (Wik, 2011). However, despite the attractiveness of the CALL approach (e.g., immediate feedback, no need for a teacher), results have been mixed; neither Neri et al (2008) or Wik (2011) found convincing improvements in production. One possibility is that this is because at least at present,

feedback cannot be responsive to a learner's individual difficulties. Talking heads are also typically highly sophisticated, showing a large amount of anatomical detail. Whilst this might be useful in making a naturalistic talking head, this may make it difficult for learners to isolate the specific information needed to help them improve their production. Finally, we took a similar approach to other studies (e.g., Herd et al., 2013), combining elements from our Perception Training program with those from our Production Training program to create our Hybrid Training condition.

TABLE I. Overview of the study design including training conditions and the pre- and post-tests completed by all participants.

Training Condition (Total participants, N = 57)	**Pre-/Post-tests**	
	Perception	**Production**
Production Training (N=16) CALVin	Task 1. Vowel identification: /b/-V-/t/ words (proportion correct)	Task 1. Recordings of /b/-V-/t/ words (a. acoustic analysis – F1/F2 midpoints, duration, b. native listener identification – proportion correct)
Perception Training (N=15)	Task 2. Category discrimination: /b/-V-/t/ words (proportion correct)	Task 2. Production of IEEE sentences (not analyzed here)
Hybrid Training (HT: N=15)	Task 3. Speech recognition in noise: IEEE sentences (speech reception threshold)	

II.METHODS

A. Participants

A total of 57 Arabic participants, aged 18-39 years old (median 27 years old), took part in the study. Eleven participants did not complete the training sessions. Of these, 2 scored over 90% at the pre-test vowel identification task and so were considered too advanced (cf. Iverson & Evans, 2009),

and 9 only completed the first session. This gave a total of 46 (18 male) participants who completed the training and all pre- and post-tests. All participants were resident in London at the time of testing but were from Arabic-speaking countries. The majority were from Saudi Arabia, with a few from other Arabic countries (2 from Egypt, 2 from Syria, 1 from Oman, 1 from Jordan, 2 from Kuwait) but all spoke a variety of Arabic that used the standard Arabic six-vowel system. Participants were recruited to have a range of experience with English (cf. Iverson & Evans, 2009). They had begun learning English at different ages (5-35 years old, median 13 years old) and had 3-69 months (median 4 years) experience of living in an English-speaking country. All participants reported no history of speech or hearing problems.

Participants were randomly assigned to one of the three training conditions: Production Training (16 participants), Perception Training (15 participants) and Hybrid Training combining training in both production and perception (15 participants).

For practical reasons, we did not aim to match participants across the different training groups based on performance on each pre-task, i.e., their abilities in speech perception and production. This meant that average performance on individual tasks varied between the training groups at pre-test (see Results). Instead, we matched participants based on their general English language skills using the written grammar section of the Oxford placement test (Allan, 1992, Wang & Treffers-Daller, 2017; cf. Iverson & Evans, 2009). The test consists of short passages with blanks in the texts: participants are required to choose the word or phrase which best fits each space from the three or four options. The test is widely used in EFL contexts and tests both L2 learners' knowledge of English vocabulary (e.g., word meanings, synonyms and antonyms) and grammatical knowledge (e.g., tense, passive voice). General language skills have been shown to be an important factor in L2 comprehension (e.g., Wang & Treffers-Daller, 2017). We hypothesized that those who had better general language skills might benefit more from training and so groups were matched on this

measure; there was a range of scores within each training group, but performance did not differ significantly between the 3 training groups, $p > 0.05$. Scores ranged from 42%-98% (median 62%) for the Production Training group; 38% - 92% (median 60%) for the Perception Training group and 38% - 72% (median 58%) for the Hybrid Training group.

In addition, 10 Standard Southern British English (SSBE) speakers (4 male and 6 female) participated in the study. They were 18-40 years old (median 21 years old), recruited from the UCL Psychology pool, and from the south of England. These participants identified Arabic learners' vowels and rated sentences for accent. Only vowel identification data is presented here. They also recorded the same /b/-V-/t/ words that were produced and identified by Arabic learners to give normative data.

Arabic participants volunteered and were given a small gift to thank them for their participation. SSBE participants were paid for their participation.

B. Apparatus

The pre-and post-tests were conducted in a quiet room with stimuli played over headphones (Sennheiser 555) via a laptop at a user-controlled comfortable level. The same laptop was used to collect responses via an experimental interface. Recordings were made using a digital audio recorder (Zoom H2 Handy Recorder) at 44,100 kHz, 16-bit resolution. All perceptual training (Perceptual and Hybrid Training conditions) was completed by participants in their own time. Participants in these conditions all used the UCL Vowel Trainer (Iverson & Evans, 2009). Participants provided their own laptops which they brought to the first training session, and the training software was installed onto their machines. The training software was password protected and on completion of each training session created password-protected log files that participants could not access. This meant that participants could not change the settings and that the researcher could verify that participants had finished the training. Production Training (Production Training and Hybrid

Training conditions) was completed with an instructor (the first author) with the aid of CALVin. Each session took 45 minutes and took place in quiet rooms.

C. Stimuli

1. Training

CALVin design. The CALVin interface was designed using a Graphical User Interface (GUI). CALVin was designed to be used as a training tool to support the teaching of English vowels. It was designed to be accessible to a range of language learners and therefore assumes no knowledge of phonetics. Briefly, the interface enables learners to; listen to and view animations (animated mid-sagittal section of a stylized talking head) of each vowel, view step-wise instructions about how to produce a given vowel, contrast recordings of different vowels within and/or between pre-defined vowel clusters, and to record their own voice so that they can compare their own production of a given vowel with that of a native speaker. Mid-sagittal section animations were intended to be accurate approximations rather than faithful physiological representations, and were based on existing descriptions of vowel production (e.g., Ladefoged, 2001); vowel production is highly variable and we hypothesized that these idealized animations would be beneficial for learners (see Alshangiti, 2015 for full details of animation process). For ease of navigation, vowels were grouped into 5 different clusters: high/front vowels: (/i: ɪ e/), open (/æ ʌ ɒ/), central/low back (/ɜ: ɑ: ɔ:/), back (/u: aʊ əʊ/) and closing diphthongs (/eɪ aɪ/). These were selected based on our previous research (Evans & Alshangiti, 2018; Iverson & Evans, 2007) and were expected to be highly confusable for many L2 learners, including Arabic learners of English.

Recordings of English words and isolated vowels were made by a male monolingual SSBE speaker. The speaker recorded three types of stimuli: keywords in a /h/-V-/d/ context, example words, and isolated vowels. The /h/-V-/d/ words included all monophthongs (/i: e ɜ: ɑ: ɔ: æ ʌ ɒ u:/) and four diphthongs (/aʊ əʊ eɪ aɪ/). The speaker recorded two example words for each vowel

(28 words: 2 examples for 14 vowels), and an example of each vowel in isolation. Example words had a CVC, CCVC or CVCC structure, e.g., *back, blouse, forks*. The words were selected to be familiar to L2 learners and were orthographically unambiguous. To ensure that the isolated vowels were as naturalistic as possible, the speaker recorded each isolated vowel after a keyword. The speaker recorded 3 repetitions of each keyword followed by the isolated vowel, all with a falling intonation contour, and the best recording was chosen for use in the experiments. Finally, all stimuli were band-pass filtered (60-20000 Hz with a smoothing factor of 10), downsampled to 22050 Hz and then saved into individual wav files before being embedded in the training software.

UCL Vowel Trainer. The recordings of the training words were made by five speakers of British English, (3 female and 2 male). The vowels were divided into four clusters: /iː/, /ɪ/, /aɪ/, /eɪ/ (e.g., *feel, fill, file, fail*); /ɛ/, /ɑː/, /æ/, /ʌ/ (e.g., *pet, part, pat, putt*); /ɒ/, /əʊ/, /ɔː/ (e.g., *was, woes, wars*); and /uː/, /aʊ/, /ɜː/ (e.g., *shoot, shout, shirt*). The clusters were based on the results of a hierarchical cluster analysis on previous English vowel identification data from L2 English speakers (Iverson & Evans, 2007) and were similar to those used in production training. There were 10 sets of minimal pair words for each of these clusters, giving a total of 140 words.

2.Pre- and post-tests

a.Vowel identification and Category Discrimination tasks. The stimuli were the same as in Iverson et al. (2012). These consisted of natural recordings of English /b/-V-/t/ (/iː ɛ ɜː æ ɑː ɒ ʌ ɔː uː eɪ aɪ aʊ əʊ/) words made by 10 SSBE speakers (5 male, 5 female). English vowels that would create non-words in the /b/-V-/t/ context (e.g., /ʊ/) were not included in the study. None of these words and speakers were used in the training corpus, such that all pre- and post-tests measured generalization to new stimuli and speakers.

b. Speech recognition in noise. The stimuli were recordings of the phonetically balanced IEEE Harvard sentences (Rothauser, 1969). There are 72 lists of 10 sentences, and each sentence contains

5 key words that are identified by the listener, e.g., "Glue the sheet to the dark blue background" (keywords underlined). The stimuli were taken from an existing recording made at UCL by a male SSBE speaker in a sound attenuated room. The speech was mixed with white noise (S. Rosen, UCL); the noise level was fixed to 71dBA, and the level of the speech was varied adaptively. Stimuli were played using a computer sound card, and participants listened over headphones in a quiet room.

D.Procedure

1.Training

All training sessions lasted approximately 45mins. The number and the duration of training sessions were the same across all training types: participants completed 5 training sessions, 1 session a day, with all 5 sessions completed over 1-2 weeks.

a. *Production Training.* All training sessions were completed with an instructor (first author), a native Arabic speaker who is highly fluent in English. Before starting the first session, participants completed a 10-minute practice session to familiarise them with the software and the relationship between the different positions of tongue, jaw and lips and the resulting vowel sound. Every effort was made to avoid technical language.

Each session followed the same structure, with all 14 English vowels (10 monophthongs, 4 diphthongs) trained in each session. At the beginning of each session, participants were trained on all 5 clusters for 10 minutes, spending more time on the vowels/contrasts they found most difficult. Clusters were worked through in the following order; high/front, open, central/low-back, back and closing diphthongs. Participants were then trained on one cluster for 20mins. They began by clicking on a keyword and then using the animation and step-by-step instructions with help from the instructor to practice producing the vowel, first in isolation and then in the example words. They then recorded themselves producing the isolated vowel, keyword, and the example words, played back their recordings, and compared them to the native speaker's production. All participants were

trained on clusters in the same order; high/front, open, central/low-back, back and closing diphthongs. For the final 10 minutes, participants compared the cluster trained that day to the other 4 clusters, starting from the diphthongs cluster and ending with the high/front cluster (i.e., the reverse of the first 10 minutes). This procedure ensured that all participants were trained on all vowels at the beginning and end, while allowing some of the training to be customised to fit the needs of each individual subject, matching the structure of the perceptual training (Iverson & Evans, 2009). All training was completed in English.

During the training sessions, participants received audio (recordings of their own production) and visual (looking at themselves in a hand mirror) feedback, as well as feedback from the instructor.

b. *Perception Training.* There was an initial 14-trial session. This was completed after the pre-tests so that participants. Each session consisted of 225 trials of vowel identification with feedback and lasted about 45 minutes. There was a different speaker in each session, as is typical of high variability phonetic training procedure (e.g., Logan et al., 1991). The procedure was the same as described in Iverson and Evans (2009).

c. *Hybrid Training.* This consisted of one session of Production (CALVin) that took approximately 45 minutes and covered all vowels, preceded by a practice session of 10 minutes, and 4 sessions of perception training (UCL Vowel Trainer). All testing and production training sessions were carried out 1-to-1 by the first author so further increasing the number of production training sessions would have been challenging. All participants began with production training, and then completed the 4 perception training sessions. The reason for this fixed order was largely practical. Further, recruiting participants for multi-session studies is challenging and we wanted to minimize the number of times participants needed to travel for an in-person session. We did not want the post-tests to be conducted immediately after a training session and so this meant that the in-person production training session had to take place during the first session after participants had

completed the pre-tests and after the perception training software had been installed on their laptop. The sessions followed the same procedure as described above.

2.Pre/post tests

Participants completed four pre- and post-tests in a fixed order, (i) vowel identification, (ii) category discrimination, (iii) speech recognition in noise, and (iv) English vowel and sentence production. These were chosen to enable us to assess changes in vowel categorization (vowel identification, category discrimination) and whether or not this led to improvements on a more real-world measure of speech perception (speech recognition in noise) or transferred to production.

a.*Vowel identification.* Participants heard natural recordings of English /b/-V-/t/ words. On each trial, they heard a word and then gave a closed-set identification response (all 14 words as response options). To give their response, participants mouse-clicked on the button which listed the stimulus word (e.g., *bout*) as well as a common English word (e.g., *house*). They received no feedback and were not able to replay the stimulus. There were six repetitions of the 14 vowels for a total of 84 trials. As in Iverson et al. (2012), the speakers were randomly mixed (i.e., all 10 were mixed within the same block) to make the task more equivalent to the category discrimination task, which requires having stimuli from different talkers.

b.*Category discrimination.* Participants heard three English /b/-V-/t/ words on each trial, spoken by three different speakers; two words were the same and one was different. Participants were asked to judge which of the three words was different (i.e., an oddity task). Participants received no feedback and were not able to replay the stimuli. There were eleven pairs of words, and each pair was played six times. Within each pair (e.g., /ɪ/-/e/), the order of presentation was counterbalanced such that half the trials were presented with /ɪ/ as the odd stimulus, and half with /e/ as the odd stimulus, with the odd stimulus played first, second, or third. The vowel pairs were: /ɪ/-/e/, /ɒ/-/ʌ/, /eɪ/-/aɪ/, /aʊ/-/əʊ/, /ɑː/-/ɔː/, /ɜː/-/ɑː/, /uː/-/əʊ/, /iː/-/e/, /uː/-/aʊ/, /ɜː/-/ɔː/, /iː/-/ɪ/. These pairs were

selected based on previous English vowel identification by Arabic speakers (Evans and Alshangiti, 2018). As in Iverson et al. (2012), the most confusable vowel pairs were selected in descending order until each of the 14 stimulus vowels appeared at least once.

c. Speech recognition in noise. We used an adaptive speech-in-noise task to estimate participants' Speech Reception Threshold. Participants heard a sentence and were asked to verbally repeat what they heard; the experimenter logged the number of correctly identified keywords. There were five keywords in each sentence, and sentences were not repeated. Each participant completed two blocks of sentences at the pre- and post-tests, selected at random from a total of 710 sentences (71 lists of 10 sentences). The first list was used as a practice session. Each block had a maximum number of 20 trials, giving a total of 40 sentences. Participants' noise threshold was found using a modified Levitt procedure (Baker and Rosen, 2001).

d.English production. All participants recorded the 14 /b/-V-/t/ words 3 times as well as the first ten sentences (List 1) of the IEEE sentences. Recordings of /b/-V-/t/ words from the pre- and post-test were analysed acoustically and were also given to 10 SSBE listeners for identification judgments, following the same procedure as in the vowel identification described above. The sentences are not analyzed here.

For the acoustic analysis, only the monophthongs were analysed. The clearest two repetitions (i.e., no hesitation, lip-smack) were chosen for acoustic analysis giving a total of 1100 analysed tokens. All measurements were made in Praat (Boersma & Weenink, 2013). The formant frequencies were measured from the midpoint where the formant frequencies were most stable. All duration measurements were taken from the beginning of the F2 transitions to the end of the F2 transitions. All F1 and F2 raw values were checked for any value 2 standard deviation outside the range, and these measurements were hand corrected, as necessary. To enable comparison of male and female data, F1 and F2 were normalised using Lobanov's z-score transformation (Lobanov, 1971) which

has been shown to be the best in factoring out physiological differences whilst preserving other sources of variation (Adank et al., 2004).

III. RESULTS

All analyses were conducted in the R statistics package (R Core Team, 2022). Either a general linear mixed effects model (continuous data; Speech Recognition in Noise, Vowel Production) using the package *lme*4 (Bates, Maechler, Bolker, and Walker, 2014) or logistic mixed effects model (categorical data; L2 vowel identification, Category Discrimination, SSBE Vowel Identification) was fitted with Test (pre-, post-), Training group (Production Training, Perception Training, Hybrid Training) and their interactions with proficiency as fixed factors and proficiency as a covariate. Random intercepts of participant and vowel/vowel pair were included as random factors, and the Holm correction (Holm, 1979) was used for pairwise comparisons. Model outputs are given in Appendix C.

A. Perception

1. Vowel Identification

Fig.1 shows vowel identification accuracy plotted by proficiency, as measured using a grammar test, for each training condition at the pre- and post-test. A logistic mixed effects model showed a significant effect of Test, χ^2 (1) = 131.54, p <.0001. A pairwise comparison using the Holm method indicated a significant change in participants' performance from pre- to post-test, β =-0.58, *SE*= 0.0502, z=-11.62, p<.001. This was modulated by Proficiency and Training group. There was a significant effect of Proficiency, χ^2 (1) = 9.56, p <.01, and a significant interaction between Proficiency and Training group, χ^2 (2) = 6.29, p <.05; participants with higher proficiency scores scored better overall, with higher proficiency participants in the Hybrid group performing best, β =0.104, *SE*= 0.045, z=2.3, p<.05. There was no significant effect of Training group, but there was a significant interaction between Training group and Test, χ^2 (2) = 43.93, p <.0001, participants in the

Perception Training group performed significantly better at the post-test, β = 1.77, *SE*= 0.51, z= 3.46, p <.001, see Fig.1. The effect of the three-way interaction between Proficiency, Test and Group was not significant.

2. Category discrimination

Fig. 1 shows the category discrimination accuracy at the pre- and post-test for each Training group. Overall, performance on this task was high for all Training groups, and there appeared to be little change in discrimination performance after training. A logistic mixed effects model showed no effects of Training group, and no significant interaction of Training group and Test but there was a significant effect of Test, χ^2 (1) = 27.204, p <.0001; a pairwise comparison using Holm method showed a significant difference from pre- to post-test, β = -0.347, *SE*= 0.066, z=-5.22, p<.001, indicating a small but

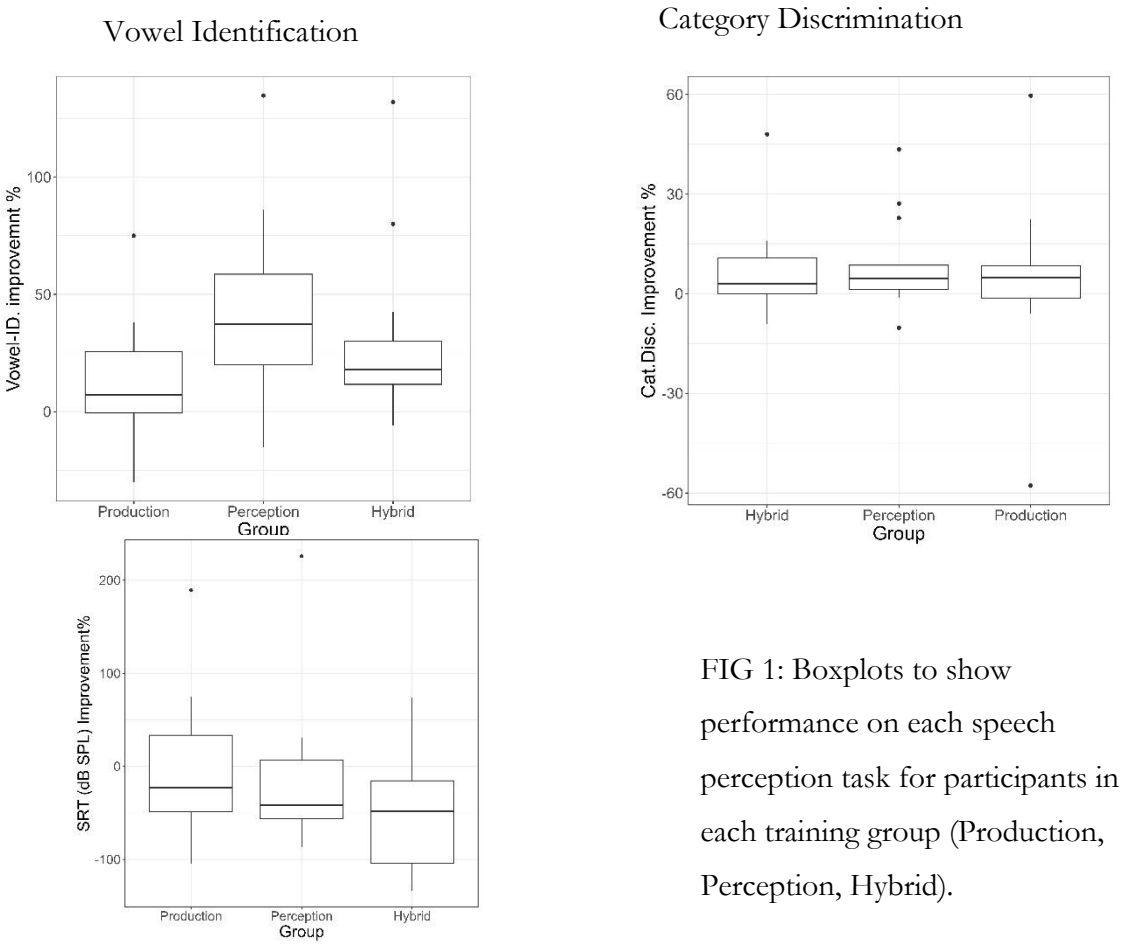

FIG 1: Boxplots to show performance on each speech perception task for participants in each training group (Production, Perception, Hybrid).

significant improvement in discrimination accuracy. Visual inspection of the data indicated that this effect was most likely driven by improvements in discrimination of /ɪ/-/e/, /ɑ:/-/ɒ/, /ɑ:/-/ʌ/, and /æ/-/ʌ/.

3. **Speech recognition in noise**

As displayed in Fig. 1, all learners with higher proficiency scores, regardless of training group, appeared to improve more in their performance on this task after training. A linear mixed model was built to examine any potential changes in the speech ratio threshold (SRT) scores. The main effect of Test was significant, χ^2 (1) = 18.09, p <.0001, and a pairwise comparison showed a significant improvement in speech recognition in noise from pre- to post-test, β =3.55, *SE*= 0.864, p<.001. There was also a significant effect of Proficiency, χ^2 (1) = 4.63, p <.05, and a significant interaction of Proficiency and Test, χ^2 (1) = 5.31, p <.05; participants with higher proficiency scores improved more after training, β =-0.39, *SE*= 0.19, p<.05, regardless of training type. However, there was no significant effect of Training group or interaction of Training group and test.

B.Speech production

1.Acoustic Analysis of /b/-V-/t/ words (monophthongs only)

Vowels from the /b/-V-/t/ words at the pre- and the post tests were extracted using Praat (Boersma & Weenink, 2013). Outliers were manually checked and hand-corrected where necessary (e.g., because of measurement error). The Mahalanobis Distance (MD: see Riverin-Coutlée et al., 2022) between the SSBE vowel production and that of the non-native speakers in the three training groups was calculated based on F1 & F2 measurements at the midpoint. MDs were used instead of Euclidean distance because as well as considering the centroid location, they also take into account

the spread and orientation of the reference distribution (see Riverin-Coutlée et al., 2022 for a detailed explanation). As such, this means that they are better able to account for the natural variability in speech production (cf. Kartushina et al., 2015). An MD of "0" indicates that the token is at the mean of the target space; this increases as the token moves further away.

All analyses were run in R (R Core Team, 2022), with MD used as the dependent variable in a linear mixed effects model. The MD values were fitted in a linear mixed model with training group, test and their interactions included as fixed factors and proficiency as covariate. Random intercepts of vowel and participants were included as random factors.

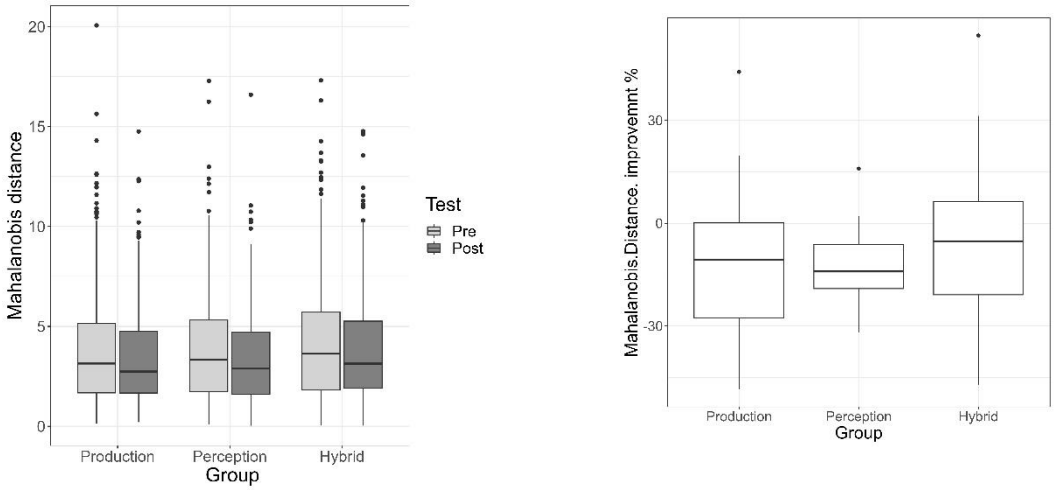

Fig. 2 displays vowel accuracy (MD) for participants in each training group, plotted according to proficiency. Mean MD values for each vowel for each training group are shown in Appendix A. There was no significant main effect of Training group, p>.05. However, there was a significant main effect of Test, χ^2 (1)=15.69, p<.0001, and a pairwise comparison using Holm's method showed

FIG.2. Boxplots to show vowel production accuracy for participants in each Training group (Production, Perception, Hybrid). Vowel accuracy is indexed by Mahalanobis Distance to SSBE speakers' vowel production: the lower the number, the closer production is to that of the SSBE speakers.

that overall, vowels were significantly closer to the target SSBE production at the post-test, β=0.419, *SE*=0.106, *t*=3.96, p<.001. There was no significant effect of the interaction between Group and Test or between Proficiency and Test.

2. Vowel Identification by SSBE listeners

/b/-V-/t/ recordings. As shown in Fig. 3, there appeared to be little overall improvement in participants' vowel intelligibility after training. A generalized linear mixed effects model was built for identification data based on the correct/incorrect binomial responses. However, there was a significant main effect of Test, χ^2 (1) = 98.33, p < .0001, and a pairwise comparison using Holm's method showed a significant change from pre- to post-test, β =-0.426, *SE*=0.0428, z=-9.96, p <.001 (median at pre-test=.64, SD = .146, and at post-test=.71, SD=.174). There was no significant effect of Training group, p >.05 but there was a significant three-way interaction between Proficiency, Training group, and Test, χ^2 (2) = 17.81, p < .001, a pairwise comparison using Holm's method showed that participants who had higher (29) grammar scores improved their performance at the post- test in the production group, β =-0.285, *SE*=0.073, z=-3.88, p<.01, in the Perception group, β =-0.463, *SE*=0.076, z=-6.08, p<.01, and in the Hybrid group, β =-0.478, *SE*=0.08, z=-5.93, p<.001.

In order to investigate whether particular vowels were more or less intelligible than others, confusion matrices for pre- and post-tests were calculated (see Appendix B). We did not conduct statistical analysis of this data, but informal inspection of the confusion matrices showed that there was notable improvement from pre- to post-test in /ɪ/, /e/ and /ɔ:/. The improvement also tended to be greater in the Production Training and Hybrid Training groups than the Perception Training group: the amount of improvement for the Production Training group was 19% for /ɪ/, 21% for /e/, and 26% for /ɔ:/; for the Hybrid Training group it was 16% for /ɪ/, 27 % for /e/and 9% for /ɔ:/; but for the Perception Training group it was 4% for /ɪ/, 13% for/e/and 19% for /ɔ:/.

Learners in the Production Training and Hybrid Training groups tended to improve more in *bit* and *bet* than those in the Perception Training group, though it is important to note that the Perception Training group were more intelligible in their production of *bet* at the pre-test than those in either the Production Training or Hybrid Training groups (Production Training = 48%,

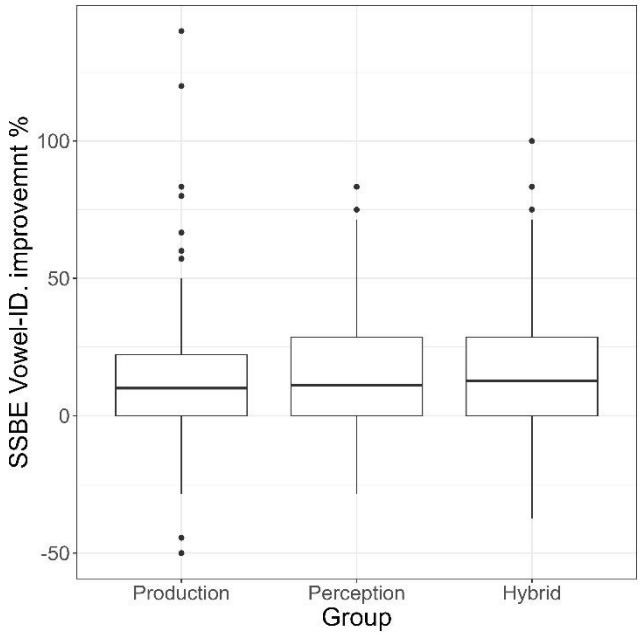

FIG.3. Boxplots to show SSBE the improvement of listeners' identification of vowels produced by speakers in each Training group (Production, Perception, Hybrid).

Perception Training = 69% Hybrid Training = 44%). This may indicate that training modified participants' production for some vowels. The word *bought* was not well identified at either the pre- or post-test for any training group (Pre-test: Production Training = 9%, Perception Training = 6%, Hybrid Training = 13%; Post-test, Production Training = 35%, Perception Training = 22%, Hybrid Training= 25%), but there was some improvement in intelligibility for all groups.

GENERAL DISCUSSION

The main findings are summarized in Table 2. All participants improved in their perception and production of English vowels after training. As expected, higher proficiency learners performed better overall but how proficiency interacted with the amount of learning depended on the task and type of training. In perception, higher proficiency learners improved more than lower proficiency learners in a sentence in noise task, and performed better overall in vowel identification. All participants were more intelligible after training, but higher proficiency learners were more intelligible than lower proficiency learners overall. Informal observation of the data revealed that key vowels that are known to be difficult for Arabic learners showed the most improvement but that this was largely limited to participants who had received explicit pronunciation training (Production & Hybrid Training groups). There was an overall improvement in production accuracy (Mahalanobis Distances), but again this was driven by improvements in higher proficiency learners in the Production Training group. There was some indication that training effects were linked to training type; learners in the Perception Training group improved significantly more in vowel identification, but not category discrimination, than did those in the Production or Hybrid Training groups, with those in the Hybrid Training group (4/5 training sessions completed using the UCL Vowel Trainer) also showing larger gains in identification than the Production Training group (though this didn't reach significance). Likewise, in production, the Production Training group improved the most in production accuracy and in vowel intelligibility (though driven by higher proficiency participants).

TABLE II. Summary of key findings

	Overall improvement?	**Effect of Proficiency?**	**Effects of Training type?**
Perception			
1. Vowel Identification	Yes	Yes: higher proficiency participants performed better overall	Yes/No: No overall effect of training type, but interaction of Training group & Test – Perception Training group performed best at post-test.
2. Category discrimination task	Yes	No	No
3.Speech recognition in noise	Yes	Yes: higher proficiency participants performed better overall and improved more than lower proficiency participants	No
Production			
1. Acoustic Analysis (monophthongs: Mahalanobis Distance)	Yes	Yes/No: No overall effect but 3-way interaction of Proficiency, Training group and Test - higher proficiency participants performed best overall.	Yes/No: No overall effect of training type, interaction of Training group & Test – Production group closer to SSBE target at post-test.
2. Vowel intelligibility	Yes	Yes/No: No overall effect but 3-way interaction of Proficiency, Training group and Test - higher proficiency participants in the Production Training group performed best overall.	No, but inspection of the data showed greater improvement for key vowels for the Production & Hybrid Training groups than the Perception Training group.

These findings are to some extent consistent with previous research showing that HVPT is particularly effective in improving perception (e.g., Logan et al. 1991; Iverson & Evans, 2009) and that production training is effective in improving production (e.g., Hattori, 2010), but that although improvements as a result of training in one modality may transfer to another (all participants showed improvement), they do not benefit the other to the same extent.

One possible reason for this finding may lie in the relationship between the tests used to measure potential improvement and the task or tasks used in training (cf. Herd et al., 2013). Like all HVPT paradigms, our Perception Training, uses an identification task with corrective feedback, with potential improvement measured using a vowel identification task, the same as the one used in training but without feedback. It is perhaps not surprising then, that our learners in our Perception Training condition improved the most in vowel identification; learners likely become more efficient and more accurate at identifying the target words and mapping them to their underlying representations. However, at least in the short-term, this may not lead to changes in underlying representations (cf., Iverson & Evans, 2009) and in turn, this may limit improvement in other, different perception tests and transfer to production. Similarly, in Production Training, participants were trained in production of vowels in isolation and in monosyllabic words, with participants tested in their production of monosyllabic words and sentences. Although there was an overall improvement in production of monosyllabic words from pre- to post-test for all participants, this was driven by higher proficiency participants in the Production Training group.

One reason why the relationship between the training and test task might be crucial is because learners may need explicit training that directs or re-directs their attention to certain acoustic cues (cf. Francis et al., 2000). In the current study, learners who completed perceptual training (Perception & Hybrid Training groups) had their attention directed to be better at identifying phonemes but not necessarily to become better at producing them. Likewise, learners in the Production Training group had training that directed them to produce native-like phonemes, but not necessarily to perceive them more accurately. Training L2 participants on production might have led to surface changes in production that resulted in more native-like production but may not have led to changes in underlying category representations which would transfer to perception (cf. Francis et al., 2000; Francis et al., 2008). Such explicit training and feedback may be particularly important for

lower proficiency learners. There is some evidence to support this in our results: higher proficiency learners appeared to benefit more generally from training indicating that they were able to generalize learning even when they did not receive explicit feedback in training in that modality.

Although it may be tempting to conclude that training is modality specific, the finding that production training did not lead to any clear improvement in vowel identification was somewhat surprising. The design of the articulatory training meant that learners listened to examples of keywords, high frequency real words, and isolated vowels, as well as their own and the instructor's examples. This meant that participants in this training modality were perceiving speech as well as receiving articulatory instructions and so we had expected that learners might improve in both production and perception. One possibility is that any production-based learning within the five sessions did not yield robust enough L2 category learning for that knowledge to be transferred to the other speech modality (i.e., perception). This could have been because participants in this training group were not exposed to as much variability: Production Training included exposure to their own speech, 1 female (instructor, first author) and 1 male talker.

Another possibility is that production may have disrupted the learning process in speech perception, in particular if perceptual representations were still fragile. Recall that our Hybrid Training group did not improve as much in vowel identification as our Perception Training group, even though four of their five training sessions were perceptual training sessions and only one session focussed on production. This differs from Herd et al. (2013) who found that participants who received either only production training or combination training (production & perception) improved as much as those in the perception training condition in perceptual identification. One possibility is that this is because of differences in the design of our hybrid training: whereas Herd et al. (2013) interleaved their production and perception training sessions, we had an initial production training session followed by perception training. It could be that the initial production training

impaired learning of new contrasts, not because of exposure to their own inaccurate production, but because of by increasing the complexity of learning (cf. Baese-Berk & Samuel, 2016, 2022). Indeed, Baese-Berk and Samuel (2016) suggest that it is only when the items are fully learned that production might aid perception. This finding is also consistent with work that has suggested that targets for articulation are defined by perceptual representations and that if these are not well-defined, production is highly variable or inaccurate (Tourville & Guenther, 2011).

Why did some vowels improve in terms of production accuracy in the Production Training group but not others? Even HP learners in the Production Training group who showed significant improvements in production accuracy after training did not change their production of the /uː/ vowel or the /ʊ-ɔː/ contrast. In SSBE and other accents of British English, /uː/ is produced as a high central rounded vowel [ʉ]. It is possible that participants were unable to hear the difference between SSBE centralized [ʉ], and Arabic [u], and instead assimilated this vowel to their native category [u] (PAM: Best, 1995; Best & Tyler, 2007). This in turn, might have prevented learning in production. This would be consistent with an explanation that production training inevitably involves perception (i.e., learning to relate new motor patterns to perceive differences between sounds) whereas the reverse need not be true. That is, if learners cannot not hear the difference between the two L2 categories, they are less likely to be able to produce them as different categories, and thus in order to train them to produce such vowels they necessarily need to be able to hear the difference between them (cf. Flege & Bohn, 2021). However, another possibility is that participants could hear the difference but were not motivated to change their pronunciation of this vowel. Although Arabic /u/ differs from the SSBE variant, this does not cause confusion with any other English vowel. Perhaps then, learning this kind of allophonic variation is not important for L2 learners, given that the aim is to be understood. Indeed, SSBE listeners also identified Arabic speakers' /uː/ accurately even if it is acoustically distant from English [ʉ].

Another possible reason for the small overall amount of improvement in production is that there was too much input during the 5 sessions for learners to adapt all the vowels. As a result, learners may have focused their attention on vowels they found particularly difficult (i.e., /ɪ, e, ɔ:/). It is possible that with more training sessions, they may have been able to change more aspects of their production. Another possibility is that, in contrast to perception, training on a smaller number of vowels might be more beneficial for production (e.g., Kartushina et al., 2015). The richness of the stimuli was initially included not only to facilitate comparison between production and perception training but because in perception, training a whole set of vowels has been shown to be more beneficial than training on a subset (Nishi and Kewley-Port, 2007). However, we know that in perception training studies in which perception training has also led to improvements in production, production of obstruents improves more than vowels and other sonorants, perhaps because these are produced with the articulators in open approximation so lack the somatosensory feedback that facilitates the learning of consonants (e.g., Herd et al., 2013; see Sakai & Moorman, 2018 for a review). Thus, whilst whole-set training might be more beneficial in perception training, training on a sub-set containing difficult contrasts may be more beneficial for training production and production of vowels in particular.

Lastly, proficiency played a role in our results. It was not surprising that higher participants performed best overall. However, initial proficiency also affected whether or not participants benefitted from training and whether they were able to generalize what they had learned to new tasks. In perception, improvement in a more real-world task of speech perception, sentence recognition in noise, proficiency not training type affected improvement. Although all learners improved in their performance in speech in noise, higher proficiency learners improved more than lower proficiency learners regardless of training type. In our study, proficiency was determined by performance on a written comprehension test that tested grammatical and lexical knowledge. One

possibility is that a certain level of grammatical and lexical knowledge is necessary to apply learning on isolated sounds and words to real-world contexts. Likewise, in production, it was higher proficiency learners who benefitted the most from Production training, with higher proficiency participants in the Production training group improving the most improvement in vowel production accuracy and in terms of vowel intelligibility. One possibility is that lower proficiency learners would have benefitted from production training on a small number of isolated contrasts. Our production training involved explicit pronunciation training as well as listening and repeating example words, so involved perception. It could be that a certain level of proficiency is needed in order to manage the cognitive load involved in training that involves both production and perception of multiple contrasts, and therefore to benefit from training on the vowel system as a whole.

In conclusion, the results from this study indicate that all 3 training types lead to improvements in L2 vowel perception and production (cf. Herd et al., 2013 for consonants), but that how training in one modality transfers to the other is complex (cf. Nagle & Baese-Berk, 2022). Whilst these findings provide support for the hypothesis that perception and production co-evolve – all participants improved from pre- to post-test to some extent – it is also consistent with the view that the correspondence between the two domains is not always perfect (Flege & Bohn, 2021: 30) and improvements in one domain may not always lead to changes in another. What improvements are observed as a result of training may also be task-driven (performance on a speech in noise task was affected predominantly by proficiency rather than by training type) or dependent on the contrast being learned (participants continued to use their L1 /u:/ likely because it was highly intelligible despite it being acoustically distant from the English vowel). This implies that whilst perception and production may share the same underlying representations, the way in which they are mapped to tasks of perception and production is likely task and context dependent.

ACKNOWLEDGMENTS

The authors are grateful to Mike Coleman for his help in building CALVin, and to King Abdulaziz University, Saudi Arabia, for funding to the first author.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Ethics Approval

The work presented in this article was approved by the UCL Ethics Committee. All participants gave informed consent before participating in any part of the study.

Data Availability

The data that support the findings of this study are available from the corresponding author upon request.

APPENDICES

APPENDIX A

Production accuracy (Mean MD, standard error in brackets) for each of the ten monophthongs in the pre- and post-tests for each training type.

Vowel	Production Training		Perception Training		Hybrid Training	
	Pre-test	**Post-test**	**Pre-test**	**Post-test**	**Pre-test**	**Post-test**
ɑ:	2.56 (0.20)	2.52 (0.32)	3.35 (0.30)	2.26 (0.29)	3.62(0.31)	3.17 (0.36)
æ	2.81 (0.32)	2.39 (0.26)	3.64 (0.33)	3.42 (0.29)	3.51(0.42)	3.09 (0.34)
i:	1.76 (0.25)	1.35 (0.14)	1.70 (0.31)	1.64 (0.14)	1.80 (0.22)	2.36 (0.19)
ɜ:	3.92 ((0.36)	3.79 (0.40)	3.91 (0.41)	3.95 (0.36)	3.89 (0.37)	4.48 (0.39)
e	8.46 (0.46	5.80 (0.52)	7.53 (0.64)	6.08 (0.49)	7.89 (0.40)	6.41 (0.48)
ɪ	3.28 (0.34)	2.64 (0.25)	2.93 (0.33)	2.76 (0.33)	3.27 (0.42)	2.29 (0.26)
u	5.09 (0.26)	4.59 (0.31)	5.13 (0.15)	4.79 (0.14)	5.16 (0.13)	4.71 (0.23)
ɒ	5.95 (0.51)	5.65 (0.72)	6.49 (0.69)	4.85 (0.50)	4.64 (0.84)	5.04 (0.60)
ɔ:	6.13 (0.68)	6.80 (0.75)	4.01 (0.46)	5.73 (0.89)	5.92 (0.83)	4.52 (0.55)
ʌ	2.82 (0.34)	2.70 (0.29)	3.05 (0.23)	1.90 (0.24)	2.73 (0.23)	2.45 (0.25)

APPENDIX B

Confusion matrix showing SSBE listeners' identification of participants' post-test production of /e, ɪ, ɔ:/ (percent correct), split by Training group (Production, Perception, Hybrid). Values are rounded to the nearest whole number.

		Response													
Group	Stimulus	eɪ	ɑ:	æ	i:	ɜ:	e	ɪ	aɪ	əʊ	u:	ɒ	ɔ:	aʊ	ʌ
Production	e	2	1	8	1	1	48	40	0	0	0	0	0	0	1
	ɪ	0	0	4	5	0	36	42	14	0	0	0	0	0	0
	ɔ:	0	3	1	0	0	0	0	0	68	0	4	9	14	2
Perception	e	2	0	4	3	3	69	19	0	0	0	0	0	0	0
	ɪ	1	0	1	5	0	45	35	13	0	0	0	0	0	0
	ɔ:	2	1	1	0	0	0	0	0	66	1	1	6	18	4
Hybrid	e	1	0	3	6	1	44	43	1	0	0	0	0	0	1
	ɪ	0	1	0	11	1	41	39	7	0	0	0	0	0	0
	ɔ:	0	1	0	0	0	0	0	7	49	2	0	13	27	1

Confusion matrix showing SSBE listeners' identification of participants' post-test production of /e, ɪ, ɔ:/ (percent correct), split by Training group (Production, Perception, Hybrid). Values are rounded to the nearest whole number.

		Response													
Group	Stimulus	eɪ	ɑ:	æ	i:	ɜ:	e	ɪ	aɪ	əʊ	u:	ɒ	ɔ:	aʊ	ʌ
Production	e	0	0	13	6	3	69	4	0	1	0	0	0	0	4
	ɪ	0	0	1	11	0	21	61	6	0	0	0	0	0	0
	ɔ:	0	1	0	0	1	0	0	0	31	0	3	35	22	7
Perception	e	1	0	9	0	2	82	5	1	0	0	0	0	0	0
	ɪ	1	0	2	0	0	51	39	7	0	0	0	0	0	0
	ɔ:	0	0	0	0	1	0	0	0	45	8	1	25	20	0
Hybrid	e	1	0	5	0	1	71	21	0	0	0	0	0	0	1
	ɪ	0	0	3	3	1	29	55	7	0	0	0	0	0	2
	ɔ:	0	0	0	0	0	0	0	0	59	3	7	22	8	1

APPENDIX C

Appendix C: Model outputs

Factors contrasts using sum code (used in all tasks)

```
#for Test
VID2$Test<-factor(VID2$Test, level=c("Pre", "Post"))
VID2$Test <- C(VID2$Test, sum)
Contrasts_Test <- contrasts(VID2$Test)
Contrasts_Test
  [,1]
pre    1
post  -1
#for group
VID2$Group<-factor(VID2$Group, level=c("Production", "Perception", "Hybrid"))
VID2$Group <- C(VID2$Group, sum)
Contrasts_Group <- contrasts(VID2$Group)
Contrasts_Group
  [,1] [,2]
Production    1    0
Perception    0    1
Hybrid       -1   -1
```

a. Vowel Identification accuracy

```
model1<-glmer(score~Proficiency*Group*Test+(1|Participant)+ (1|Vowel), data=VID, family = binomial, nAGQ=0 )
```

Analysis of Deviance Table (Type II Wald chisquare tests)
Response: score

	Chisq	Df	Pr(>Chisq)
Proficiency	9.5623	1	0.001986 **
Group	1.7165	2	0.423913
Test	131.5462	2	< 2.2e-16 ***
Proficiency:Group	6.2980	2	0.042895 *
Proficiency:Test	0.0480	1	0.826647
Group:Test	43.9382	2	2.877e-10 ***
Proficiency:Group:Test	5.3408	2	0.069223

Fixed effects:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.458261	0.513171	-2.842	0.004488 **
Proficiency	0.060448	0.016317	3.705	0.000212 ***
Group1	-1.732369	0.645370	1.989	0.046703 *
Group2	0.439302	0.613226	0.716	0.473758
Test1	-0.204712	0.127151	-1.610	0.107399
Proficiency:Group1	-0.049156	0.020953	-2.346	0.018977 *
Proficiency:Group2	-0.012049	0.020447	-0.589	0.555686
Proficiency:Test1	-0.003298	0.004319	-0.764	0.445124
Group1:Test1	0.308519	0.167055	1.847	0.064774 .
Group2:Test1	-0.579520	0.162453	-3.567	0.000361 ***
Proficiency:Group1:Test1	-0.003508	0.005484	-0.640	0.522426
Proficiency:Group2:Test1	0.012601	0.005502	2.290	0.022013 *

Post-hoc tests:

```
test_emm <- emmeans(model1, ~ Test)
pairs(test_emm, adjust = "Holm")
 contrast   estimate     SE  df z.ratio p.value
 Pre - Post   -0.584 0.0502 Inf -11.624  <.0001

group_emm <- emmeans(model1, ~ Group)
pairs(group_emm, adjust = "Holm")
 contrast              estimate    SE  df z.ratio p.value
 Production - Perception  -0.2022 0.266 Inf  -0.760  1.0000
 Production - Hybrid      -0.0921 0.266 Inf  -0.346  1.0000
 Perception - Hybrid       0.1100 0.271 Inf   0.406  1.0000
```

For the interactions:

```
roup_by_test_emm <- emmeans(model1, ~ Test|Group)
>   pairs(group_by_test_emm, adjust = "Holm")
Group = Production:
 contrast   estimate     SE  df z.ratio p.value
 Pre - Post   -0.218 0.0824 Inf  -2.651  0.0080

Group = Perception:
```

contrast estimate SE df z.ratio p.value
Pre - Post -1.037 0.0908 Inf -11.416 <.0001

Group = Hybrid:
contrast estimate SE df z.ratio p.value
Pre - Post -0.561 0.0892 Inf -6.291 <.0001

3-way onteractions

group_by_test_emm <- emmeans(model1, ~ Test|Group|Proficiency)
> pairs(group_by_test_emm, adjust = "Holm")
Group = Production, Proficiency = 29.8:
contrast estimate SE df z.ratio p.value
Pre - Post -0.197 0.0843 Inf -2.343 0.0191

Group = Perception, Proficiency = 29.8:
contrast estimate SE df z.ratio p.value
Pre - Post -1.015 0.0917 Inf -11.065 <.0001

Group = Hybrid, Proficiency = 29.8:
contrast estimate SE df z.ratio p.value
Pre - Post -0.605 0.0958 Inf -6.315 <.0001

b. Category Discrimination

model2<-glmer(Score~Proficiency*Group*Test+(1|Participant)+ (1|Pair), data=AXB3, family = binomial, nAGQ=0)

Analysis of Deviance Table (Type II Wald chisquare tests)
Response: Score

	Chisq	Df	Pr(>Chisq)
Proficiency	1.2164	1	0.2701
Group	0.7044	2	0.7031
Test	27.2047	1	1.83e-07 ***
Proficiency:Group	0.7277	2	0.6950
Proficiency:Test	1.1369	1	0.2863
Group:Test	0.6712	2	0.7149
Proficiency:Group:Test	0.9501	2	0.6219

test_emm <- emmeans(model2, ~ Test)
pairs(test_emm, adjust = "Holm")
contrast estimate SE df z.ratio p.value
Pre - Post -0.347 0.0664 Inf -5.223 <.0001

c. Speech recognition in noise

model3<-lmer(SRT~Proficiency*Group*Test+(1|Participant),data=SpNz)

Analysis of Deviance Table (Type II Wald chisquare tests)
Response: SRT

	Chisq	Df	Pr(>Chisq)
Proficiency	4.6334	1	0.03135 *
Group	3.1097	2	0.21123

Test	18.0974	1	2.099e-05 ***
Proficiency:Group	0.8371	2	0.65801
Proficiency:Test	5.3160	1	0.02113 *
Group:Test	0.5497	2	0.75967
Proficiency:Group:Test	3.3047	2	0.19160

Fixed Effects:

	Estimate	**Std.Error**	**df**	**t value**	**Pr(>|t|)**
(Intercept)	20.091565	4.453620	40.196868	4.511	5.49e-05 ***
Proficiency	-0.339812	0.148904	40.076235	-2.282	0.0279 *
Group1	-3.678218	5.903147	40.015454	-0.623	0.5368
Group2	0.501811	5.633553	40.834773	0.089	0.9295
Test1	-3.594793	2.096970	39.536947	-1.714	0.0943 .
Proficiency:Group1	0.171759	0.191347	39.957948	0.898	0.3748
Proficiency:Group2	0.013006	0.187099	40.539424	0.070	0.9449
Proficiency:Test1	0.184032	0.069884	39.422827	2.633	0.0120 *
Group1:Test1	-10.13030	8.57653	40.00725	-1.181	0.2445
Group2:Test1	-0.518133	2.697011	40.116628	-0.192	0.8486
Proficiency:Group1:Test1	-0.161881	0.089516	39.309470	-1.808	0.0782 .
Proficiency:Group2:Test1	0.009251	0.088894	39.853088	0.104	0.9176

```
model3<-lmer(SRT~Test+(1|Participant),data=SpNz)
test_emm <- emmeans(model3, ~ Test)
pairs(test_emm, adjust = "Holm")
 contrast   estimate    SE   df t.ratio p.value
 Pre - Post     3.55 0.864 44.3   4.102  0.0002
```

d. Vowel Production: Mahalanobis Distance

```
model4<-lmer( Mahalanobis.Distance ~ Proficiency*Group*Test+(1 |Participant) +
(1|Vowel),data=DS2, REML=T)
anova(model4)
```

Analysis of Deviance Table (Type II Wald chisquare tests)
Response: Mahalanobis.Distance

	Chisq	**Df**	**Pr(>Chisq)**
Proficiency	1.5559	1	0.2123
Group	3.1140	2	0.2108
Test	15.6911	1	7.457e-05 ***
Proficiency:Group	0.1226	2	0.9406
Proficiency:Test	0.0007	1	0.9792
Group:Test	1.0316	2	0.5970
Proficiency:Group:Test	2.1651	2	0.3387

Fixed effects:

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	4.244e+00	6.996e-01	2.910e+01	6.067	1.31e-06 ***
Proficiency	-1.568e-02	1.661e-02	4.000e+01	-0.944	0.351
Group1	1.471e-01	6.589e-01	4.000e+01	0.223	0.824
Group2	-1.008e-01	6.249e-01	4.000e+01	-0.161	0.873
Test1	1.901e-01	2.675e-01	1.779e+03	0.711	0.477
Proficiency:Group1	-5.962e-03	2.137e-02	4.000e+01	-0.279	0.782
Proficiency:Group2	-3.556e-03	2.080e-02	4.000e+01	-0.171	0.865
Proficiency:TestPost	0.02478	0.03274	75.08229	0.757	0.4514
Group1:Test1	-3.827e-01	3.550e-01	1.779e+03	-1.078	0.281
Group2:Test1	3.788e-01	3.367e-01	1.779e+03	1.125	0.261
Proficiency:Group1:Test1	1.380e-02	1.151e-02	1.779e+03	1.198	0.231
Proficiency:Group2:Test1	-1.133e-02	1.121e-02	1.779e+03	-1.011	0.312

model4<-lmer(mahalanobis~Test+(1|Participant)+ (1|Vowel), data=DS2)
test_emm <- emmeans(model4, ~ Test)
pairs(test_emm, adjust = "Holm")
contrast estimate SE df t.ratio p.value
Pre - Post 0.419 0.106 1784 3.963 0.0001

e. Vowel identification by SSBE listeners

model5<-glmer(Score~Proficiency*Group*Test+(1|Speaker)+ (1|Vowel), data=VIDN, family = binomial, nAGQ=0)

Analysis of Deviance Table (Type II Wald chisquare tests)
Response: Score

	Chisq	Df	Pr(>Chisq)
Proficiency	0.9206	1	0.3373134
Group	1.3014	2	0.5216877
Test	98.3397	1	< 2.2e-16 ***
Proficiency:Group	4.7562	2	0.0927282 .
Proficiency:Test	4.5246	1	0.0334110 *
Group:Test	2.7397	2	0.2541471
Proficiency:Group:Test	17.8174	2	0.0001352 ***

Fixed Effects:

	Estimate Std.	Error	z value	Pr(>|z|)
(Intercept)	-0.037641	0.679018	-0.055	0.95579
Proficiency	0.027588	0.015390	1.667	0.095504 .
Group1	1.103272	0.604541	1.825	0.068005 .
Group2	0.225350	0.580857	0.388	0.698045
Test1	0.876111	0.283957	3.085	0.00203 **
Proficiency:Group1	-0.041349	0.019522	-2.118	0.034170 *
Proficiency:Group2	-0.005000	0.019331	-0.259	0.795921

Proficiency:Test1	-0.007286	0.003645	-1.999	0.045599 *
Group1:Test1	0.504230	0.141900	3.553	0.000380 ***
Group2:Test1	-0.450413	0.135129	-3.333	0.000859 ***
Proficiency:Group1:Test1	-0.014863	0.004651	-3.196	0.001394 **
Proficiency:Group2:Test1	0.014219	0.004557	3.121	0.001805 **

```
test_emm <- emmeans(model5, ~ Test)
pairs(test_emm, adjust = "Holm")
 contrast   estimate     SE  df z.ratio p.value
 Pre - Post   -0.426 0.0428 Inf  -9.962  <.0001
```

REFERENCES

Adank, P., Smits, R., & Van Hout, R. (2004). A comparison of vowel normalization procedures for language variation research. The Journal of the Acoustical Society of America, 116(5), 3099-3107.

Allan, D. (1992). *Oxford Placement Tests 1.* Oxford University Press, Oxford, UK.

Alshangiti, W. (2015). *Speech production and perception in adult Arabic learners of English: A comparative study of the role of production and perception training in the acquisition of British English vowels.* Ph.D thesis, University College London, UK.

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.

Baese-Berk, M. M., & Samuel, A. G. (2022). Just give it time: Differential effects of disruption and delay on perceptual learning. *Attention, Perception & Psychophysics* 84: 960-980.

Baker, R. J., Rosen, S. (2001). Evaluation of maximum-likelihood threshold estimation with tone-in-noise masking. British Journal of Audiology 35(1), 43-52.

Bates D, Maechler, M., Bolker B., & Walker, S. (2014). lme4: Linear Mixed-Effects Models Using Eigen and S4. R package version 1.1-7, URL http://CRAN.Rproject.org/package=lme4

Best, C. T. (1995). A Direct Realist View of Cross-Language Speech Perception. In *Speech perception and linguistic experience: Issues in cross-language research*, ed. W. Strange, pp. 171-204. York Press.

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

Boersma, P., & Weenink, D. (2013). Praat: doing phonetics by computer [Computer program]. Version 5.3.51.

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.

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.

Flege, J. E., & Bohn, O. S. (2021). The revised speech learning model (SLM-r). In *Second language speech learning: Theoretical and empirical progress*, ed. R. Wayland, pp. 3-83. Cambridge University Press.

Flege, J.E., Takagi, N. & Mann, V. (1995). Japanese adults can learn to produce English /r/ and /l/ accurately. Language & Speech 38: 25-55.

Francis, A. L., Baldwin, K., & Nusbaum, H. C. (2000). Effects of training on attention to acoustic cues. Perception & Psychophysics 62(8), 1668-1680.

Francis, A. L., Kaganovich, N., & Driscoll-Huber, C. (2008). Cue-specific effects of categorization training on the relative weighting of acoustic cues to consonant voicing in English. Journal of the Acoustical Society of America 124(2), 1234-1251.

Ghaffarvand Mokari, P., & Werner, S. (2019). On the Role of Cognitive Abilities in Second Language Vowel Learning. *Language and Speech*, *62*(2), 260-280. https://doi.org/10.1177/0023830918764517

Hanulíková, A., Dediu, D., Fang, Z., Bašnaková, J., & Huettig, F. (2012). Individual differences in the acquisition of a complex L2 phonology: A training study. Language Learning 62: 79-109.

Hattori, K. (2010). *Perception and production of English/r/-/l/ by adult Japanese speakers*. Ph.D thesis, University College London, UK.

Hattori, K., & Iverson, P. (2009). English /r/-/l/ category assimilation by Japanese adults: individual differences and the link to identification accuracy. Journal of the Acoustical Society of America 125(1), 469–79.

Herd, W., Jongman, A. & Sereno, J. (2013). Perceptual and production training of intervocalic /d, ɾ, r/ in American English learners of Spanish. Journal of the Acoustical Society of America 133(6),4247- 4255

Holes, C. (2004). *Modern Arabic: Structures, functions, and varieties*. Georgetown University Press.

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70. http://www.jstor.org/stable/4615733

Hubert, M., Debruyne, M., & Rousseeuw, P. J. (2018). Minimum covariance determinant and extensions. *Wiley Interdisciplinary Reviews: Computational Statistics*, *10*(3), e1421. https://doi.org/10.1002/wics.1421

Jarrah, M. A. S. (1993), The Phonology of Medina Hijazi Arabic: A Non-Linear Analysis. Ph.D. thesis, University of Essex.

Inceoglu. S. (2016). Effects of perceptual training on L2 vowel perception and production. *Applied Psycholinguistics* 37(5): 1175-1199.

Iverson, P., & Evans, B. G. (2007). Learning English vowels with different first-language vowel systems: Perception of formant targets, formant movement, and duration. Journal of the Acoustical Society of America 122(5), 2842-2854.

Iverson, P., Evans, B. G. (2009). Learning English vowels with different first-language vowel systems II: Auditory training for native Spanish and German speakers. Journal of the Acoustical Society of America 126(2), 866–77.

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y. I., Kettermann, A., & Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. Cognition 87(1), B47-B57.

Iverson, P., Pinet, M., Evans, B. G. (2012). Auditory training for experienced and inexperienced second-language learners: Native French speakers learning English vowels. Applied Psycholinguistics 33(1), 145-160.

Ladefoged, P. (2001). *A Course in Phonetics*. Heinle & Heinle.

Lambacher, S. G., Martens, W. L., Kakehi, K., Marasinghe, C. A., & Molholt, G. (2005). The effects of identification training on the identification and production of American English vowels by native speakers of Japanese. Applied Psycholinguistics 26(2): 227–247.

Leach, L. & Samuel, A.G. (2007). Lexical Configuration and lexical engagement: When adults learn new words. *Cognitive Psychology* 55(4): 306-353.

Lengeris, A., & Hazan, V. (2010). The effect of native vowel processing ability and frequency discrimination acuity on the phonetic training of English vowels for native speakers of Greek. Journal of the Acoustical Society of America 128(6): 3757–68.

Lobanov, B. M. (1971). Classification of Russian vowels spoken by different speakers. Journal of the Acoustical Society of America 49(2B), 606-608.

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English/r/and/l: A first report. The Journal of the Acoustical Society of America 89(2), 874-886.

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. Journal of the Acoustical Society of America 138(2),, 817-832.

Massaro, D. W., & Light, J. (2003). Read my tongue movements: bimodal learning to perceive and produce non-native speech/r/and/l. In *Eighth European Conference on Speech Communication and Technology*.

Nagle, C. & Baese-Berk, M. (2022). Advancing the state of the art in L2 speech perception-production research: Revisiting theoretical assumptions and methodological practices. *Studies in Second Language Acquisition* 44: 580-605.

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(5), 393-408.

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.

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from << https://www.R-project.org/>>.

Riverin-Coutlée, J., Roy, J.-P., & Gubian, M. (2022). Using Mahalanobis Distances to Investigate Second Dialect Acquisition: A Study on Quebec French. Language and Speech 66(2), 291–321.

Rothauser, E. H. (1969). IEEE recommended practice for speech quality measurements. IEEE Trans. on Audio and Electroacoustics 17, 225-246.

Saito, K. & Plonsky, L. (2019). Effects of Second Language Pronunciation Teaching Revisited: A Proposed Measurement Framework and Meta-Analysis. *Language Learning* 69(3): 652-708.

Sakai, M. & Moorman, C. (2018). Can perception training improvethe production of second language phonemes? A meta-analytic review of 25 years of perception training research. *Applied Psycholinguistics* 39: 187-224.

Shafiro, V., Levy, E. S., Khamis-Dakwar, R., & Kharkhurin, A. (2013). Perceptual Confusions of American-English Vowels and Consonants by Native Arabic Bilinguals. Language and Speech 56(2), 145-161

Shinohara, Y. & Iverson. P (2018). High variability identification and discrimination training for Japanese speakers learning English /r/–/l/. Journal of Phonetics 65: 242-251.

Thomson, R. (2012). Improving L2 Listeners' Perception of English Vowels: A Computer-Mediated Approach. *Language Learning* 62(4): 1231-1258.

Tourville, H. A., & Guenther, F. H. (2011). The DIVA model: A neural theory of speech acquisition and production. Language and Cognitive Processing 26(7), 952–981.

Trofimovich, P., & Gatbonton, E. (2006). Repetition and Focus on Form in Processing L2 Spanish Words: Implications for Pronunciation Instruction. *Modern Language Journal* 90(4): 519–535.

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

Wells, J.C. (1982). *Accents of English*. Cambridge University Press.

Wik, P. (2011). *The Virtual Language Teacher: Models and applications for language learning using embodied conversational agents*. Doctoral dissertation, KTH Royal Institute of Technology