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

Wafaa Alshangiti

A thesis submitted in fulfilment of requirements for the degree of Doctor of Philosophy

To

Department of Speech, Hearing and Phonetic Sciences

Division of Psychology and Language Sciences

Faculty of Brain Sciences

University College London (UCL)

2015

Declaration

I confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Wafaa Alshangiti

Abstract

This thesis presents the results of four studies that investigated the perception and production of English by Saudi Arabic learners. Additionally, the thesis sought to investigate the role of different types of training, production- or perception-based, in learning, with the aim of understanding how training in different domains contributes to second language acquisition.

A preliminary study (Study 1) investigated problematic phonemic contrasts for Arabic speakers, confirming that accuracy in perception and production depends on the similarity between L1 and L2 phonemes.

Study 2 investigated the specificity of second language phonetic training by comparing the effect of three training programmes on the acquisition of British English vowels. Saudi Arabic learners were randomly assigned to one of three training programmes; Production Training (PT), High Variability Phonetic Training (HVPT), and a Hybrid Training Program (HTP). They completed a battery of tests before and after training. All participants improved after training, but improvements were largely domain-specific; production training led to improvements in production but not perception, whilst perception training led to improvements in perception but not production. Participants in the HTP showed improvements in both production and perception, indicating that only a small amount of training in production appears to be necessary to effect changes in production. Additionally, improvement on particular tasks appeared to be linked to initial L2 proficiency, and learning in perception and production was retained (Study 3) and production training appeared to be more beneficial for participants who were trained in a non-immersion setting (Study 4).

In brief, the results suggest that L2 learners improve in both perception and production if training explicitly trains these domains. Production training was beneficial not only for L2 learners in an L2-speaking country, but also in non-immersion settings. Overall, these results suggest that a hybrid training programme would be most beneficial for L2 learners.

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Bronwen G Evans and Valerie Hazan for their support and guidance throughout the course of my PhD. Bronwen has always been there when needed, she made it easy for me to drop by her office asking for advice, and managed to find time in her busy schedule to answer my questions and send me useful comments on my work. She encouraged me to be more critical and to develop my ideas. Valerie kindly gave me useful feedback on several stages of the course, and I learnt a lot from her discussions in her lab meetings, where she found the time to comment on and discuss my work. I am also grateful to Paul Iverson for providing me with the vowel trainer that he and Bronwen developed.

I would also like to thank a number of people who have helped me during my research; Mike Coleman for his patience in giving me feedback about the animation used in CALVin, and for his efforts in putting things together in an interface program. His help make it easier to develop the vowel trainer for production, Stephen Nevard for his willingness to help when things went wrong in the lab, and David Cushing for resolving any technical problems.

During my PhD I have received help and support in both my personal and professional learning from my PhD colleagues, and I am really lucky to have had the chance to be with a group of really nice and supportive people. They share knowledge, give tutorials of the area of their expertise, and make learning a new programme or skill easier. Special thanks to José Joaquín Atria for his help with Praat scripts, Yasna Pereira Reyes, Sonia Granlund and Tim Schoof for their academic and personal support, Yasuaki Shinohara for his great help with "R", Mark Wibrow for English Language advice, Dorothea Hackman, and to all folks in the PhD room who make it a wonderful environment to work; Georgina Oliver, Lucy Carroll, Mauricio Figueroa-Candia, Albert Lee, Kurt Steinmetzger, Nada Al-Sari, Csaba Redey-Nagy, Dong-Jin Shin, Emma Brint, Daniel Kennedy-Higgins, Louise Stringer, Gisela Tome Lourido , Jieun Song, Faith, Cristiane Hsu, Hao Liu and Yue Zhang. I would also like to thank

all the anonymous participants in the thesis. Special thanks to Mohammed Alqarni for his help with participant recruitment.

My PhD is funded by King Abdul-Aziz University, and I am grateful for their financial support without which my PhD wouldn't have been possible.

Finally, my children have been a great support to me even with their age, they have been my source of strength, and they know that mum is studying, and sometimes cannot go with them on school trips or assemblies. I am grateful for Rabeya Ullah and her daughters for their help throughout my study, and for Emily Sollof, Sally Taylor, and Kate Pepper for their help and support. Thanks to my dad for his encouragement, to my father-in-law for his support, and of course to Ahmed for his financial support throughout my studies.

Table of Content

List of Figures

List of Tables

Chapter 1 Introduction

1.1 FOREWORD

The increased influence of English as a lingua franca with the quick pace of globalisation has brought to the fore research into second language learning. Such research has examined learning from many different perspectives, but since conversation forms a fundamental means for communication compared to other language skills (e.g., reading & writing), one of most frequently studied aspects of learning a second language (L2) has been whether and how learners are able to acquire the skills to speak and understand (i.e., produce and perceive) L2 phonemes accurately in order to be understood and to understand others (e.g., Morley, 1991).

One of the emerging themes in such research has been that whilst learners can acquire the phonemes of a non-native language late in life (i.e., post critical period) to some degree, successful production and perception of the L2 is affected by the relationship between their native language (L1) and L2 phonemes (e.g., Iverson et al., 2003). For example, Japanese learners of English find it hard to identify and produce English /r/ and /l/ (e.g., Goto, 1971), a contrast which does not exist in their native language. Similar effects have been found for vowels; Spanish learners find the /i/-/ɪ/ contrast, a contrast which does not exist in their native language, difficult to perceive and produce (e.g., Flege & MacKay, 2004), often producing words such as 'ship' with a long vowel, such that it sounds closer to 'sheep', and confusing the contrast in perception. It is thus easy to see how these difficulties can lead to misunderstandings and mishearings in conversational settings, particularly in everyday conversation where listening conditions may be difficult.

Understanding what problems a learner faces in acquiring the phonemes of their L2 successfully, and thus, being able to design appropriate and successful training programmes, necessarily entails understanding the relationship between the L1 and L2 in question. This thesis thus presents, as its starting point, a study of the perception and production of English phonemes by Arabic learners of English from a wide range of proficiency levels (Study 1). To my knowledge, no study has investigated Arabic speakers' acquisition of British English as second language (though see Shafiro et al., 2012 for a study of bilingual Arabic-English speakers), yet Arabic speakers represent

a large and influential group of L2 English users, especially given their strong links to the UK and USA through business and increasingly, education.

Additionally, the thesis sought to investigate the efficacy of different training techniques and in so doing, add to our understanding of the nature of the link between speech production and perception. The relationship between speech perception and production has been a long-standing focus in speech science. Several theories of speech perception have suggested strong links between speech perception and production (e.g. Liberman et.al, 1985), arguing that both processes share common underlying representations - a view supported by brain imagining studies (e.g., Wilson et al., 2004) which show that areas of the brain involved in speech production are activated during listening.

However, despite such links between production and perception, studies of L2 learning have not consistently demonstrated that perceptual training leads to improvements in production and vice versa. Previous studies have shown that perceptual training techniques, specifically High Variability Phonetic Training (HVPT), are beneficial for improving the perception of difficult L2 phonemic contrasts (e.g., Logan et al., 1991), and some have found that this training generalizes to production, at least for some learners. For example, Bradlow et al. (1997) showed that after intensive perceptual training for the /r/-/l/ contrast (45 sessions over 3-4 weeks), Japanese speakers improved in their perception and were also able to transfer this learning to the production domain. Similar effects have also been found for vowel perception and production (see Lambacher et al., 2005).

By contrast, others have found little or no relationship between perceptual learning and production, suggesting that perception and production operate somewhat independently. For example, Hattori (2009) trained Japanese speakers on English /r/-/l/ production over 10 one-to-one sessions using a multi-faceted approach that used explicit feedback from the instructor, real-time spectrograms, and feedback with synthesised versions of their own productions. Hattori found that after intensive production training, Japanese speakers improved their production to become more native-like, but that training did not improve their perception of English /r/-/l/ at all.

The second part of this thesis thus presents the results of a training study (Study 2) that aimed to examine whether the type of training affected learning of English vowels

by Arabic learners of English. Learners were assigned to one of three vowel training programs: Production Training (PT), High Variability Phonetic Training (HVPT) and a Hybrid Training program (HTP) which included both production and perception training. Each training program aimed to give learners the same amount of training, but differed in its focus. A battery of pre- and post-tests assessed improvements in production and perception. Also of interest was whether production training was retained as well as perceptual training (Study 3), and whether learning was affected by participants' learning environment (i.e., immersion vs. non-immersion; Study 4).

In brief, the aim of the experiments represented in this thesis was twofold. First, to investigate the problematic phonemic contrasts for Arabic learners of English. Second, to further examine the link between speech perception and production in relation to training type.

1.2 OVERVIEW

Chapter 2 reviews previous work on language-specific perception and production models for L1 and L2 speech perception, and the factors that affect second language acquisition. The chapter also reviews previous studies that have investigated phonetic laboratory training including techniques such as HVPT, and various different approaches to production training. The review aims to give an overview of the phonemic contrasts that Arabic learners of English are likely to find challenging, and to make predictions about how training one speech domain might affect the improvement of the other, as well as whether combining the two speech domains in a training program might enhance L2 learning. The chapter concludes with a summary of the aims and hypotheses of the thesis.

Chapter 3 presents the results of Study 1, which investigated the perception and production of English phonemes by Arabic learners of English. The study comprises two separate experiments conducted with the same participants; vowel identification in quiet and in noise, and consonant identification in quiet and in noise. Section 34 describes the results from the experimental tasks, and section 3.5 discusses the implications of these results in light of current theories of L2 acquisition.

Chapter 4 introduces and provides the motivation for the training study (Study 2). The chapter reviews previous studies that have investigated the link between speech perception and production and then goes on to present the training study in Chapter 5.

Chapter 5 presents the overall design and the methodology used in the experiments, the results of the training study and their implications for existing theories of L2 learning as well as their relevance for the debate surrounding the link between production and perception.

Chapter 6 presents the results of Study 3 which investigates the effect of training type on retention of learning. A subset of the participants who took part in the training study (Study 2) completed the same battery of pre-/post-tests, 6 months after training. Section 6.4 presents the results and Section 6.5 discusses their implications.

Chapter 7 presents the results of Study 4, which investigated the effect of the learning environment on training. In particular, I was interested in whether L2 learners would respond differently to the production-based training programme if they were regularly in contact with native speakers and used English as a mean of daily interaction (i.e., an immersion setting), or if they had very few opportunities to interact with native speakers (i.e., a non-immersion setting). A group of participants in Saudi Arabia completed the PT programme and their results were compared with those who were tested in London. Section 7.4 presents the results of the study and Section 7.5 discusses the results and the implications for the use of production training in L2 teaching.

Finally, Chapter 8 begins by summarizing and discussing the main findings of Studies 1-4. The chapter goes on to provide a discussion of the findings within the context of existing models of speech perception and production and second language learning, before addressing the implications for current and future work.

Chapter 2 Literature Review

2.1 Language-specific speech perception and production

During the early months of life, infants appear to be sensitive to the phonetic properties that differentiate phonetic segments in any language (Eimas et al., 1971; Miller and Eimas, 1983). However this ability seems to diminish as infants reach six months of age (e.g., Kuhl et al., 1992; Polka & Werker, 1994; Bosch, & Sebastian-Galles, 2003). The linguistic experience in infants' L1 gradually modifies their sensitivities from being language-general to being more language-specific, a process which happens earlier for vowels than for consonants (Kuhl et al., 1992; Werker & Tees, 1984). Thus by the end of the first year of life, infants share, with adults, similar perceptual limitations for non-native contrasts (see Jusczyk, 1997, for a review). These perceptual modifications reflect the influence of language experience on initial phonetic sensitivities.

The transition of speech perception from language-general to language specific has been investigated by a number of studies for both vowel and consonant phoneme discrimination. For instance, Werker and Tees (1983, 1984) found a language-specific decline in sensitivity around the age of 10-12 months old. They tested 10 English-speaking adults, 5 native Thompson (native-Indian- spoken in south central British Columbia) speaking adults, and 3 Hindi, 2 Thompson and 12 English infants. The results demonstrated that English infants from 6-8 months were able to discriminate the English bilabial contrast /p/-/b/ as well as two non-English contrasts; the Thompson glottalized velar/uvular contrast /ʔki/-/ʔqi/ and the Hindi voiceless, unaspirated retroflex/dental contrast /ʈa/-/t̪a/. By 8-10 months, a smaller percentage could discriminate the non-native contrasts, and by 10-12 months, infants performed as poorly with non-native contrasts as the young children and adults. Similarly, Bosch and Sebastian-Galles (2003), tested 4 and 8 month-old infants from Spanish monolingual, Catalan monolingual and Spanish-Catalan bilingual environments in the perception of vowel contrasts present only in Catalan, /e/-/ɛ/. Again, younger infants were able to perceive this contrast regardless of the language exposure, a result which has been replicated in a number of other studies (e.g., Best 1995; Kuhl, 1998; Werker and Lalonde, 1988; Werker and Curtin, 2005).

Models of L1 speech perception (Best, 1994, 1995; Best and McRoberts, 2003; Kuhl, 2000; Kuhl et al, 2008; Werker and Curtin, 2005) postulate that the shift from language-general to language-specific perception is due to the individual's L1 learning experience. For example, the expanded Native Language Magnet (NLM-e; Kuhl et al., 2008) explains the shift from language-general to language-specific perception in view of acoustic similarity between L1 and L2Categories. It claims that the more L1 linguistic input infants receive, the more they shape their neural commitment to their native language (e.g., Kuhl, 2004) which, in turn, leads to distortion in their perceptual map, so-called perceptual warping. This perceptual wrapping, known as the perceptual magnet effect leads to a reduction in non-native language perceptual abilities, but facilitates L1 processing because it increases sensitivity to between-category contrasts whilst decreasing sensitivity to within-category variation.

The Perceptual Assimilation Model (PAM; Best, 1994, 1995; Best and McRoberts, 2003) explains the shift from language-general to language-specific perception in view of articulatory similarity between native and non-native segments. Best and McRoberts (2003) proposed "the articulatory organ hypothesis" which predicts that sensitivity towards non-native phonemes declines when two phonemes share the same primary articulator (i.e., within organ contrasts); that is if native and non-native phonemes share the same articulatory organs, older infants (10-12 months) will find it more difficult to discriminate the phonemes. Best and McRoberts tested infants in three isiZulu contrasts; [kʰ a]-[kˈa], [ɬ]-[ɮ], and [pu]-[ɓu]. These contrasts are each within-organ laryngeal distinctions; either involving a non-native laryngeal gesture (velar ejective) [kʰ a]-[k'a], a native laryngeal distinction in the context of a non-native supralaryngeal gesture pattern (lateral fricatives) [ɬ]-[ɮ], or laryngeal distinction that occurs but is non-contrastive in the native language (bilabial stops) [pu]-[ɓu]. The results showed that 6-8 month olds could discriminate all three isiZulu, while the 10-12 month olds failed to discriminate between these contrasts. They concluded that older infants show a decline in discriminating non-native within-organ distinction compared to 6-8 moth olds.

PRIMIR (Processing Rich Information from multi-dimensional Interactive Representation; Werker and Curtin, 2005) assumes that infants use general learning

mechanisms and filters to develop multidimensional interactive representations which allow for information grouping on the basis of similarity, and that this explains their perception shift from language-general to language specific. The information is grouped into three multidimensional planes; the General perception plane, the Word Formation plane, and the Phonemic plane.

Initially, infants are thought to process phonemes through the General perception plane which includes phonetic information. They statistically analyse the speech that they hear and then cluster this information in order to establish language-specific categories. Such categories help the formation of the Word Form plane. In this plane, infants extract sequences forming phonetic units in speech, and match such combination of phonetic units without meaning and object knowledge (i.e., concepts). PRIMIR claims that when infants later hear the same spoken word or words, referring to certain object, from different speakers in different examples and in different contexts, they statistically analyse which units match which referent, and eventually map the incoming phonetic units with the referent objects without errors. It is from this that the phonemic plane is predicted to emerge, with the phonemic categories becoming more robust as infants expand their vocabulary and learn to read. Once the phoneme plane is firmly established, phonemes should be readily utilized across a variety of tasks. PRIMIR then argues that this is why older children are more successful at accessing phonetic detail when learning novel words.

Despite the fact that infants acquire their L1 easily, adults typically find it challenging to acquire their L2, particularly late in life. Indeed, early L2 research claimed that it was difficult if not impossible for individuals past a certain age to successfully learn L2 sounds (e.g., Lenneberg, 1967; Wood and Loewenthal, 1981; Munro et al., 1996). It was hypothesized that this was due to the assumption of a critical period existence; after puberty, maturational changes in the brain were thought to affect an individual's ability to learn a new language. Such that around this age, they were unable to acquire a new language in the same way as their L1.

However, evidence from more recent studies of L2 acquisition is inconsistent with the existence of a critical period hypothesis (CPH) for language acquisition. Flege

et al. (1995) investigated English language learning in native Italian speakers who had been living in Canada for an average of 32 years. Subjects recorded five short sentences that were presented to native English speakers and rated for perceived foreign accent. Based on the CPH, one would have expected to see a sharp decline in accent ratings corresponding with the age at which language learning ability became impaired, e.g., around puberty. However, there was no discontinuity in the accent ratings: The ratings decreased systematically as participants' age of arrival (AOA) increased, resulting in a near-linear relationship between AOA and accent ratings.

If the decline in language learning cannot be attributed to a biologically delimited critical period, then how else might it be explained? One explanation is that, as previously discussed, early experience with a native language constrains subsequent language learning, such that one's native language interferes with the acquisition of non-native speech sounds (see Kuhl, 2000 for a review). These perceptual changes are believed to be the reason behind the difficulties that adults face when distinguishing non-native phonemes, depending on the degree of the conflict between L1 and L2 phonemes (Best, 1994; Flege, 1995; Harnsberger, 2001).

For instance, Iverson et al., (2003) found that language-specific perceptual processing can modify the relative salience of category acoustic variation, and that this can interfere with L2 acquisition. That is, adults' experience with their native language and their use of the acoustic cues by which they distinguish different phonemes, affects their sensitivity to L2 cues and thus, interference between L1 and L2 phonetic cues occurs. For example, when trying to discriminate /r/-/l/, Japanese learners are more sensitive to F2, an irrelevant cue for discrimination of the English contrast but which is associated with the Japanese /ɾ/, than they are to F3 onset frequency, the cue that is used by English native speakers (Iverson et al., 2003).

Likewise for vowels, McAllister et al (2002) demonstrated that learners who use duration contrastively in their L1 were better at acquiring vowel categories that differ according to duration, than were those who did not use this cue in their L1. McAllister et al (2002) compared the perception and production of Swedish vowel duration by L2 learners from different L1 backgrounds; American English, Latin

American English, and Estonian speakers and found that Estonian speakers who use duration contrastively in their L1 outperformed American speakers who use duration only as a secondary cue in their L1 (cf. Hillenbrand et al., 2000). However, American speakers outperformed the Spanish speakers who do not use duration distinctively in their L1. The authors argued that their results confirmed the salience of L1 transfer when learning L2 vowels, and led them to formulate a hypothesis suggesting that L2 features that are not phonologically contrastive in an individual's L1 are harder to perceive and produce.

It is important to note though, that other studies have indicated that individuals remain sensitive to novel acoustic features, in particular duration, even when they do not use those features in their L1. For example, Bohn (1995) tested German, Spanish and Mandarin speakers in their identification of American English vowels, and found that duration was used not only by German speakers who use this feature contrastively in their L1, but also by Spanish and Mandarin speakers who do not use it distinctively in their L1. Similarly, other studies (e.g., Flege et al, 1997; Escudero & Boersma, 2004; Escudero, 2005; Cebrian, 2006) have shown that Spanish speakers use duration as a cue to discriminate the English tense-lax contrast, /iː/-/ɪ/, a contrast which is not present in their L1. Based on his results, Bohn (1995) hypothesised that, when spectral information is not available, L2 learners use duration even if they do not use it contrastively in their L1, as it is an accessible and salient cue for vowel identity.

Despite the ability to use duration cue even if it is not used in individual's L1, it is clear that perception and production of one's L2 is largely affected by one's L1. Such that learning is harder for L2 phonemes that are similar to L1 phonemes than for L2 phonemes that are dissimilar to L1 phonemes (e.g., Flege, 1995; Best et al, 1988; Guion et al 2000; Flege et al, 2003).

Indeed, three of the most influential theories of L2 speech perception, Flege's Speech Learning Model (SLM; Flege, 1995), Best's Perceptual Assimilation Model for second language learning (PAM-L2; Best & Tyler, 2008) and the NLM-e/Perceptual Interference account (Kuhl et al., 2008; Iverson et al., 2003) attribute variability in the perception and production of non-native segments to the similarity

between L1 and L2 phonetic or phonemic categories. PAM-L2 (Best & Tyler, 2007) states that the difficulty in differentiating non-native phonemic contrasts is predictable from the basis of the relationship between the L1 and L2 phoneme inventories, such that depending on the relationship between the L1 and L2 phonology, L2 phonetic segments will be differently assimilated into existing L1 categories. The model proposes several possible patterns of assimilation which can account for the different levels of perceptual difficulties seen in L2 learners. Discrimination of two L2 phones is thought to be most difficult if both phones are assimilated equally well or poorly into the same single L1 category, a single-category pattern of assimilation. For example, both Thompson Salish ejective velar /k'/ and uvular /q'/ are likely to assimilate to English /k^h/, although both will be heard as strange or discrepant from the English standard. A two-category assimilation pattern occurs when two different non-native phones are assimilated into two different L1 categories. In this case, excellent discrimination accuracy is predicted. For example, the Hindi retroflex stop /ɖ/ is likely to assimilate to English [d], while Hindi breathy-voiced dental stop /$d^ɦ$/ may assimilate a different English phoneme category, the voiced dental fricative [ð]. However, when two L2 phonemes are assimilated into a single L1 category with one of the phonemes being a closer match to the L2 category than the other, the assimilation is categorised as a category-goodness contrast. In this case, PAM predicts that listeners will have moderate discrimination accuracy. For example, both Zulu voiceless aspirated velar /k/ and /k'/ are likely to assimilate to English /k^h/, but the former should be perceived as identical with English standard while the latter should be heard as quit discrepant from it. The Uncategorized-Categorized contrast occurs when one of a two L2 phonemes is identified with an L1 category, and the other is not assimilated to any L1 category (i.e. one L2 phoneme is categorised and the other is not categorised). In this case PAM predicts that listeners will have high discrimination accuracy. The Uncategorized-Uncategorized contrast occurs when both L2 phonemes are not assimilated to any L1 category. In this case, PAM predicts that discrimination accuracy will vary from poor to moderate depending on the proximity of the two L2 phonemes to other L1 phonemes. Besides these patterns of assimilation, Best's model predicts that if one or both of the non-native phoneme contrasts are sufficiently phonetically dissimilar from any native category, they may be classified as

unassimilable to any L1 speech sound. The perceptual discrimination of such phonemes thus depends on their phonetic or auditory similarity to each other rather than on their relationship to L1 categories. For example, the suction-produced click consonants of southern Bantu languages are unlikely to assimilate to any English phoneme categories.

Flege's Speech Learning Model (SLM; Flege, 1995, 1999, 2002) also predicts that performance with an L2 depends on the relationship between the L1 and L2 categories. The SLM proposes that the capacity for L2 learning remains intact across the life span, and that experience plays a salient role in changing the way in which the L1 and L2 phonetic subsystems interact. This model proposes that L2 segments that are phonetically similar, but not identical, to L1 categories are perceptually assimilated to those L1 categories; even after considerable experience with an L2, the perceptual representation of similar L1 and L2 phonetic segments may not be differentiated, but rather may be a compromise between L1 and L2 categories. However, the greater the distance between the perceived L2 sound and the closest L1 sound, the more likely it is that the phonetic differences between the sounds will be detected, and a new phonetic category will eventually be established (Flege, 1995). L2 categories that are perceptually distinct from any L1 category are not assimilated to L1 category, and are thus easier to learn since they fall into relatively unoccupied regions of the listener's phonological space.

Similarly, the Perceptual Interference account (Iverson et al., 2003) proposes that L2 learners use their L1 phonetic cues to perceive and produce L2 phonemes, and that this causes phonetic interference between the L1 and the target L2. This interference is thought to occur as a result of the 'mis-tuning' of the perceptual space for L2 contrasts, which can make the irrelevant acoustic variation in the L2 more salient than the critical differences in L2 phonetic cues (Iverson et al., 2003). That is, the perceptual space is optimally tuned for the L1, such that a native speaker is more sensitive to between- rather than within-category variation. This set of tunings might not be applicable to the L2, as sensitivity to meaningful contrasts in the L2 might be reduced, making accurate perception and production of L2 contrasts difficult.

Although these three models make different predictions about the origin of difficulties, they all explain these difficulties in terms of the relationship between L1 and L2 categories, rather than attributing them to for example, maturational constraints. However, L2 learners from the same L1 background can acquire L2 differently, some can be more successful than others. This suggests that there are other factors that play a role in L2 speech perception and production.

Although research has shown that there is substantial variability in L2 learners' difficulties in learning non-native categories (e.g., Lively et al, 1993), some difficulties in perceiving and producing phonemes are often not predictable from comparing native and non-native phoneme-inventories (Kohler, 1981). Some non-native sounds are easy to perceive even with no prior experience of a target language, such as L1 English listeners' perception of Farsi velar versus uvular stops (Polka, 1992), and of voicing and place contrasts in Zulu clicks (Best et al., 1988). However, contrasts which are the same as native contrasts in terms of phonological features can be difficult for L2 learners. For example, Gottfried (1984) found that American English (AE) learners of French find it difficult to discriminate between French rounded vowels /u-y/, /y-ø/, because they are more similar to AE vowels /u/ and /ʊ/. Gottfried also found that the AE learners who were experienced in French made fewer errors than the AE listeners who did not speak French, which suggests an effect of experience on identification accuracy. Similar findings have been found in other studies (e.g., Levy and Strange, 2008).

2.2 Additional factors influencing L2 speech perception & production

In addition to the finding that L1 experience affect L2 learning, L2 research has proposed a number of other factors that help explain why L2 acquisition is harder for adult learners and why some learners are more successful than others, even when they are from the same L1 background. These factors can be assigned to broader categories including factors concerned with the age of L2 learning (Flege et al, 1995), length of residence in an L2-speaking environment, duration of learning, formal education, the degree of L2 use in daily life (Piske et al, 2001), and the relative quantity and quality of input from native L2 speakers (Flege & Liu, 2001; Flege, 1999, 2002; Flege & MacKay, 2004; Jia & Aaronson, 2003; Jia, et al., 2006).

Previous studies have shown that, as far as L2 learning is concerned, 'the earlier the better' (e.g., Flege, 1995b; Flege et al, 1999; Flege, 2007). Individuals who began learning their L2 in late adolescence or early adulthood (late bilinguals) usually resemble native speakers of the L2 less than individuals who began learning their L2 in childhood (early bilinguals) do. Such early bilinguals typically have a milder foreign accent than do late bilinguals, and have been found to perceive and produce L2 vowels more like native speakers than do late bilinguals (Munro et al, 1996; Flege et al, 1999; Meador & Flege 2001). They are also better at detecting speech in noise than are late bilinguals (Meador et al., 2000).

One crucial factor in L2 learning success is the amount of L1 usage relative to that of the L2. For example, Flege (1997; see also Piske & Flege, 2001) demonstrated that native Italian speakers who used their L1 frequently had a stronger foreign accent when speaking English, their L2, than those who used their L1 infrequently. The degree of the L1 and L2 use might also explain the differences early vs. late bilinguals (L2 learners henceforth); arguably early bilinguals have had the opportunity to use both their languages for a longer period of time than individuals who learnt another language later in life. Moreover, L2 learners are more likely to have less high-quality L2 input than most early bilinguals. Another possibility is that the input that adult learners is perceiving maybe not targeted to their level (e.g., grammatical and lexical knowledge), and thus, it might be much harder for them to learn.

Length of residence (LOR) in an L2 speaking country/environment has also been shown to affect L2 learning. For example, Flege and Liu (2001) compared two groups of Chinese students with different length of residence in Canada (2 versus 7 years), and found that the students who had 7 years of residence in Canada outperformed the ones who only been in Canada for 2 years in three measure of L2 learning (identification of word final consonants, listening comprehension and grammatical sensitivity to English sentences). They concluded that LOR matters for the learners, especially for those who need to use English on a daily basis (students). They also suggested that not only LOR but also how much native-speaker input they received, affects the way that L2 learners progress in English (L2).

For instance, a child who moves to their host country early in life is usually enrolled in nursery or school where they have many opportunities to interact with their peers and teachers, often (but not always) native speakers of their L2. Thus, they are more likely to complete a relatively high level of schooling, form social relationships and possibly marry a fluent English speaker if not a native speaker. In contrast, adult L2 learners are more likely to be working in environments where they interact with speakers who share the same L1 or with other speakers from different L1 backgrounds, and less interactions with the native speakers of the L2. Also, they are likely to have less education in their L2, since they likely completed their education in their home country (Stevens, 1999).

2.3 L2 Phonetic training

Despite the fact that learning an L2 is challenging, previous research has shown that there is some flexibility in the adult system to support non-native category learning (e.g., Bradlow et al, 1997; Goudbeek et al, 2008; Iverson et al, 2005). Adults can learn to discriminate acoustic differences between non-native sounds that they may not be able to categorize linguistically, at least within experimental tasks (Werker and Tees, 1984). Moreover, some studies provide evidence to show that L2 learners can learn difficult L2 phonemic contrasts if they receive sufficient training or directed input (e.g., Logan et al., 1991; Bradlow, 1997; Iverson et al, 2005; Iverson & Evans, 2007, 2009; Hattori, 2009). These studies show that after intensive training in laboratory settings, L2 learners improve in their perception and production, that learning generalizes to new stimuli and speakers, and that it is also retained over relatively long periods of time (e.g., Iverson & Evans, 2009).

Many of these training studies have used the High Variability Phonetic Training (HVPT) paradigm. Originally developed by Logan et al. (1991), this involves participants giving identification judgments with corrective feedback on a given word or phoneme that is produced by several speakers in different phonetic contexts. For example, Logan et al. (1991) used HVPT to train native Japanese speakers to distinguish the English /r/-/l/ contrast. Participants completed 15 training sessions over a three week-period, with each session lasting 40 minutes. To assess training, they were given a battery of pre- and post-tests; /r/-/l/ identification task, and generalization

tests that consisted of novel words (not included in training) in order to measure if learning generalized to new stimuli.

The results demonstrated that identification of the /r/-/l/ contrast improved by 7.8% after training and that they could generalize the training to new stimuli and new speakers who were not included in the training. Subsequent studies have shown that training in the perceptual domain can also transfer to production. For instance, Bradlow et al (1997) investigated the effect of training Japanese speakers on perceptual identification of English /r/-/l/. They were given 45 sessions of perceptual identification training with feedback over a period of 3-4 weeks. Before and after training, they were tested in their identification of /r/-/l/ minimal pairs and were also asked to record English words that contrasted /r/ and /l/. Japanese listeners' identification improved by 16% in the post-test, and more interestingly, for some participants at least, improvements in perception led to improvements in production as well.

HVPT has also been used successfully to train L2 learners on vowel stimuli. For example, Lambacher et al., (2005) trained native Japanese speakers over 6 weeks on the identification of American English mid and low vowels /æ/, /ɑ/, / ɔ/, / ʌ/, /ɜ/. Participant completed 6 sessions of training, (each session took approximately 20 minutes), and were tested before and after training in vowel identification task. As in previous studies, the results demonstrated that Japanese speakers improved in their identification of the target vowels, and their improvement proceeded to production of the targeted vowels.

However, some cues appear to be less responsive to HVPT than others. Hirata et al. (2007) and Tajima et al. (2008) used HVPT to train English speakers on Japanese vowel length contrasts. Hirata et al. (2007) trained English speakers who completed 4 training sessions over 11-17 days. The target words were nonsense Japanese words in the context of /mVmV/ and /mVmVV/ (V=/ i, e̞, ä, o̞, ɯᵝ/, e.g., /mimi/ vs /mimiː/). Participants were assigned to one of the three training groups; slow-rate, fast-rate, and slow-fast training materials where speakers were instructed to speak as slowly as possible and as fast as possible respectively. Participants were tested before and after

training on minimal pairs embedded in a carrier sentence (e.g., /saju/-/saju:/). The results showed that perceptual abilities for English speakers generalized from trained rates to tested rates. That is, slow-fast training showed effects not only on the participants' overall scores, but also on the three rate tests including the normal rate which was not included in the training. However the overall effect of training was small; 9.1% for slow-fast training.

Similarly, Tajima et al., (2008) trained English speakers for three training sessions over five days on minimal pair identification using isolated words contrasting in vowel length, produced at normal rate, a fast rate, and a slow rate. The target vowels were; /i, e̞, ä, o̞, ɯᵝ/. Participants were tested in minimal pair identification before and after training. The results showed that training improved perception of the trained contrast types, however, the improvement did not generalize to the untrained contrasts.

The results from these studies showed that there was a small degree of improvement in perception, but unlike previous studies, learning did not generalize to new tokens. One possibility is that this is because listeners were only trained using a sub-set of the vowels (e.g. 5 out of 15 including diphthongs in Lambacher et al, 2005), or using closed-set responses (e.g., long vs short as in /kado/ versus /ka:do/, Tajima et al., (2008)).

Indeed, Nishi and Kewley-Port (2007) suggested that training individuals on a subset of vowels does not help them generalize learning to untrained vowels, and that it is more efficient to train individuals on larger set of vowels. They trained two groups of Japanese native speakers on American English vowels; one group was trained on 9 vowels (full set training group), and the other on only 3 vowels (subset training group). They found that participants in the full-set group improved in identification by 25%, and they were able to generalize learning to untrained vowels, while the sub-set group did not improve in the untrained vowels.

At least for vowels, training on a large dataset seems to be beneficial. Iverson and Evans (2007, 2009) used 5 sessions of HVPT to train Spanish and German speakers, over 1-2 weeks, on an even larger set of vowels. Learners were trained on 14 English vowels including diphthongs (e/, /ɑ:/, /æ/, /ʌ/, /i:/, /ɪ/, /aɪ/, /eɪ/, /ɒ/, /əʊ/, /ɔ:/,

/uː/, /aʊ/, /ɜː/), to increase the range of variability. To make the stimuli close to real-world communication, minimal pairs were used in multiple environments (e.g., *pet, part, pat, putt, feel, fill, file, fail, was, woes, wars, shoot, shout, shirt*) unlike other studies where they use CVC or nonsense words. Participants completed a battery of tests before and after training; English vowel identification in quiet, English vowel identification in noise, L1 vowel identification in noise, L1 assimilation, and vowel space mapping. The results showed that after 5 sessions of HVPT both Spanish and German learners improved in their vowel perception, with learning retained 4-12 months after training (Iverson & Evans, 2009). However, despite the fact that they improved in their vowel identification, their best exemplar locations did not improve. Iverson and Evans (2009) argued that this was because HVPT helped learners to apply their existing knowledge about L2 vowel categories to L2 identification more successfully, but that it did not change learners' underlying representations of these categories.

Even so, HVPT appears to be a highly successful way of improving learners' identification of difficult phonetic contrasts. Learners improve rapidly in their perception of difficult contrasts over a relatively small number of sessions, and are able to apply this learning to new speakers and new phonetic contexts. There is also some evidence that training in the perceptual domain transfers to production (e.g., Bradlow et al., 1997; Lambacher et al., 2005). Indeed, the relationship between speech perception and production has been a long-standing focus in speech science and several theories of speech perception have suggested strong links between speech perception and production. For example, Liberman et.al (1985), argu that both processes share common underlying representations - a view supported by brain imaging studies (e.g., Wilson et al., 2004) which show that areas of the brain involved in speech production are activated during listening. However, despite such links between production and perception, studies of L2 learning have not consistently demonstrated that perceptual training leads to improvements in production and vice versa.

Hattori (2009) used a production training technique to train Japanese learners on the production of English /r/-/l/. In the first session of training, Japanese learners

watched a slow motion video of the words *lens* and *wrens,* while the instructor provided an articulatory description (e.g., lip and tongue shape). Participants made several recordings of training words, non-words and received feedback on F3 from the instructor using Praat. In sessions 2-8, participants began with visual and perceptual comparisons between pronunciations, and compared them to signal-processed versions of minimal pairs (e.g., *rack-lack*), that they had recorded at the first session. Then all participants practiced producing /r/ and /l/, repeating after the instructor, while receiving feedback on F3, closure duration and transition duration. This feedback was given using a real-time spectrogram. In sessions 2-8, they were instructed to produce lengthened versions of the words so that these acoustic cues were easier to track, but by sessions 9 &10, they practiced producing the training words with natural closure and natural transition durations. After 10 intensive one-to-one sessions, Hattori found that Japanese speakers improved their production of /r/ and /l/ so that it was close to that of native speakers, yet their perception of this consonant contrast did not improve at all. Hattori concluded that although speech perception and production are somehow related, their underlying mechanisms remain independent, and learning in perception and production occur at different rates. Hattori thus concluded that learning is domain specific; that is, perceptual training largely trains perception and production largely trains production.

If training is domain specific, then a hybrid approach that combines production and perception training should lead to improvements in both production and perception. In attempt to train both speech domains, Macdonald (2011) trained English learners of French on two problematic contrast, the French /u/-/y/ (oral contrast) and the /ɑ̃/-/ɔ̃/ (nasal contrast), using different training conditions over six sessions that took place over 4-6 weeks. Participants were randomly assigned to one of the training groups; pronunciation training only (listen and repeat, with pronunciation instructions), HVPT with pronunciation training (3 sessions HVPT, and 3 sessions pronunciation), perceptual fading only ("this technique attempts to train perceptual contrast, without subject errors, by starting off with clearly discriminable stimuli which may exaggerate the normal perceptual differences or add other salient features"; personal communication, R. Macdonald, [22/09/2012 & 19/11/2014], perceptual fading with pronunciation training (3 sessions of perceptual fading, and 3 sessions of

pronunciation training), and HVPT only. There was also a control group who completed the pre- and post-tests, but received no training. Participants completed minimal pair identification task, and recorded the same minimal pairs before and after training. Additionally, a subset of participants returned after a month for retention tests [post-test and 2 generalization (familiar speaker and new words, and new speaker and new words) tests]. The results showed that pronunciation training did not improve speakers' production and there was little evidence that HVPT improved production especially for the oral vowels /u/-/y/; for this vowel contrast, no group performed higher than the control group. However, for the nasal contrast, /ɑ̃/-/ɔ̃/, all training groups outperformed the control group. Macdonald concluded that perceptual training is best for improving perception, and that removing some or all of the perceptual training has an adverse effect on learning. Examination of pronunciation data collected during training suggested a slight advantage of pronunciation and HVPT + pronunciation training, though this was small and was not examined statistically.

In summary, HVPT appears to be a highly effective way of improving the perception of difficult L2 contrasts. However, a fundamental part of L2 learning is developing the ability not just to understand these phonemic contrasts, but also to accurately produce them. Although there is some evidence to suggest that training in the perceptual domain transfers to production (e.g., Bradlow et al. 1997), other studies have shown little or no transfer. Additionally, studies that have trained production have found little evidence of transfer of learning from production to perception (e.g., Hattori, 2009; Macdonald, 2011). However, these studies are small in number, have focussed on a very limited number of contrasts and are labour-intensive, involving a large number of one-to-one-training sessions (e.g., Hattori, 2009). This thesis aimed to further investigate the relationship between training type and learning, and more broadly, to better understand the link between production and perception. Additionally, it aimed to develop a more practical approach to training pronunciation. In order to do this in our target population, it was necessary to better understand the problems that Arabic learners have in acquiring English. Consequently, the next chapter, Study 1, investigated the problematic phoneme contrasts for Arabic learners of English.

Chapter 3 Speech perception and production by adult Arabic learners of English

3.1 Introduction

As previously discussed, language-specific experience has been found to influence the perception and production of L2 phonemic contrasts by L1 learners, typically when one or both phonemes in the contrast are realised differently or do not occur in the learner's L1. The current chapter describes a study designed to investigate speech perception and production in Arabic learners of English with different proficiency levels.

Although Arabic speakers potentially represent one of the largest groups of L2 English users and in many Arabic countries English is "generally viewed in a positive way and as the language of technology, progress, and the future" (Nickerson and Camiciottoli, 2013 p. 333), little previous experimental research has investigated Arabic speakers' perception and production of English as a second language. What work there is, has generally focussed on early bilingual English-Arabic speakers (e.g., Shafiro et al., 2012). They tested early Arabic-English bilinguals (with different Arabic Dialects), and native English speakers of the English dialects spoken in the United Arab Emirates, in American English vowel identification in CVC context, and consonant identification in three vocalic contexts /ɑCɑ/, /iCi/, /uCu/. Overall vowel identification for the Arabic speakers was 70%, and 80% for the native English speakers. Consonant identification accuracy was also high for both groups; 95% for the Arabic speakers and 94% for the native speakers. Closer examination of the results showed that though their overall vowel identification was high, Arabic learners found some vowels (e.g., American English /ɑ/, /ɔ/, /æ/) more confusable than consonants. Although the pharyngealised /ɑ/ in Arabic is very similar to the English /ɑ/, Arabic /a/ very similar to /æ/, yet Arabic learners find them confusable.

Given the much smaller vowel space of Arabic, it is perhaps somewhat surprising that vowel identification performance was high. However, these participants were early bilinguals with high proficiency in English. It is thus highly likely that adult L2 learners (not early bilinguals) would have more difficulty in accurately

perceiving and producing English phonemes because of the relationship between their L1 and L2 (see Chapter 2).

However, the relationship between L1 and L2 is somewhat more complicated in Arabic than in other non-diglossic languages. As in other diglossic languages, Arabic has a high and low variety; the high variety is only used in written forms and in formal settings (i.e., classical Arabic) while the low variety is used in daily conversations (i.e., dialectal Arabic). Dialectal Arabic differs from the classical Arabic in phonology, syntax, and lexicon. Recently the term Modern Standard Arabic (MSA) has emerged referring to standard Arabic, a variety that uses standard Arabic lexicon, but preserves the phonological norms of speaker's dialect (Watson 2002). There is also much variation between low varieties from different parts of the Arab-speaking world. Since the phonemic categories in different dialects may influence listeners' category assimilation, and given the fact that other studies into Arabic phonetics (e.g., Bani-Yassin and Owens, 1987) have found that some Arabic dialects have a vowel inventory that differs from the three MSA vowel, /i/, /u/, /a/ (Newman, 2002). For example, Moroccan Arabic has five vowels /iː, ə, aː, ʊ, uː/, and Jordanian Arabic an eight vowel system /iː, i, eː, a, aː, oː, u, uː/ (Al-Tamimi, 2007). It is possible that Arabic learners' difficulties with English vowels might vary according to their dialect background. Difficulties with English consonants may also be similarly affected. For example, Iraqi speakers replace MSA /q/ and /k/ with /g/ and / ʧ / in the vernacular (Alani, 1978).

3.2 The current study

The aim of this study was to investigate the perception and production of English vowels and consonants by native Saudi Arabic, especially Hijazi Arabic, learners of English. Participants with different proficiency levels (measured by a grammar test, see page 37) were recruited, enabling investigation of possible effects of proficiency. Saudi Arabic learners are frequently exposed to English from a young age in their home country, in particular through the media, and it was hypothesized that these participants, even those considered to have little direct experience with English (e.g., by living in the UK) might perform well in phoneme identification tasks. To avoid the possibility of ceiling effects, participants completed vowel and consonant phoneme identification tasks in quiet and in noise. Native English controls also

completed the vowel and consonant identification in noise. To investigate production, Arabic participants were also recorded producing the /h/ -V- /d/ vowel stimuli and a short passage (The North Wind and the Sun; IPA Handbook, 1999). English native speakers then identified the vowels produced by Saudi speakers, and rated their speech for accentedness.

In order to provide L1 reference data, a pilot study explored the vowel and consonant variants used in Hijazi Arabic, the dialect spoken in the western region of Saudi Arabia and the area from which participants in this study were recruited. Twelve speakers (5 males) aged 19-35 years (median 27 years old) from Hijaz (N=6) and Riyadh (the central region in Saudi Arabia, (N=6) were recorded completing various different tasks that elicited Arabic in different speech styles; reciting the Quran, reading and retelling a story, naming pictures in their dialect, and completing sociolinguistic interview. The results showed that Saudi speakers used the low variant [ɡ] in informal settings for the high variant /q/, and that they used /dʒ/ in formal speech and when reciting the Qur'an, while in the less formal settings they used the low variant [ʒ]. Hijazi females also used the variant [t] more than Hijazi males, while non-Hijazi females used /θ/ more than non-Hijazi males. Surprisingly, there was no difference between Hijazi and non-Hijazi males in using the variants [t, θ]; both used /θ/ more than female (Hijazi and non-Hijazi) in their speech, but tended to use [t] less. All speakers used a similar vowel inventory to that of MSA (see Appendix 2), but tended to use more central vowels.

Based on these results, it was hypothesized that Arabic learners would perform better with English consonants than with vowels. For Arabic consonants, in addition to using the standard 28 MSA consonants (e.g., Holes, 2004), Saudi Arabic speakers also use other variants such as [ʒ, ɡ]. This will likely facilitate the accurate perception and production of English consonants which map well to the Arabic consonant inventory. However, the same cannot be said for vowels; Standard Southern British English is typically described as having 12 monophthongs, and 8 diphthongs (e.g., Cruttenden, 2014), while Arabic has 6 monophthongs (3 tense, 3 lax) and 2 diphthongs. This makes it hard to map one English vowel to one Arabic vowel, and

which will therefore likely make it more difficult for Arabic learners to perceive and produce English vowels accurately (though cf. Shafiro et al., 2012).

3.3 Methodology

3.3.1 Participants:

Twenty-six Saudi Arabic speakers (from Hijaz and Riyadh) were recruited and completed a battery of tests to assess their English production and perception. Nine native Standard Southern British English (SSBE) listeners were recruited as controls and completed a subset of the perception tasks to give normative data. These SSBE listeners also completed identification and ratings tasks for Arabic participants' English production. All participants were 18-35 years old (median 26 years), reported no speech or hearing problems and were resident in London at the time of testing. Saudi participants volunteered to take part in the study, and to thank them for volunteering they were given souvenirs with the UCL logo on. SSBE speakers were paid for their participation.

Arabic speakers were recruited to cover a range of proficiency levels and had acquired English at different ages (see appendix 3). Participants began learning English when they were 2-23 years old (median 11 years), and had 3 months-9 years of experience living in the UK (median 3 years). Proficiency was assessed using the Oxford English Grammar Test (Allan, 1992).

3.3.2 Stimuli and Apparatus

Consonant perception (Quiet and Noise)

A male monolingual SSBE speaker recorded the English consonants in VCV contexts. The speaker recorded three versions of each consonant /b, p, m, w, f, v, ϴ, ð, s, z, ʃ, ʧ, ʒ, ʤ, t, d, n, r, l, g, k, ŋ, h / in two vocalic contexts /iCi/, and /ɑCɑ/ embedded in the carrier sentence "Say __ again". The vocalic contexts were varied because this has been shown to have a great effect on the phonemic perception (cf. Strange et al., 2007). Recordings were made in sound-attenuated audio booths using a Røde NT-1A microphone connected to an Edirol UA-25 sound card and saved as uncompressed 16 bit 44100 Hz LPCM files. Each word was checked for clarity and the clearest word

was selected, down-sampled to 22050Hz and amplitude-normalised to 70 dB SPL. Stimulus sets for the consonant perception in quiet test used these selected recordings unaltered. The stimulus sets for the consonant perception in noise test were created by mixing the selected recordings with speech-shaped noise (S. Rosen, UCL) generated by a Wandel and Goltermann RG-1 noise generator at three signal-to-noise (SNR) ratios (0, -5, and -10 dB). In order to create speech in noise conditions, the root mean square (RMS) amplitude of the stimulus and noise were determined and scaled to fit the SNR condition. They were then combined through addition at the three SNRs using an automated script in Praat (Boersma & Weenink, 2005). Finally, all stimuli files were equalized for intensity at 70 dB SPL. The three noise levels were chosen to vary the difficulty of recognizing the words, and the order of the noise level was randomised.

3.3.2.1 Vowel perception (Quiet and Noise)

The vowel stimuli were recorded by the same male SSBE speaker used for the consonant stimuli. Three versions of 17 vowels covering mostly the whole vowel space were recorded; /iː, ɪ, e, æ, ɑː, ɒ, ɔː, uː, ʊ, ʌ, ɜː, eɪ, aɪ, aʊ, əʊ, eə, ɔɪ/. Vowels were produced in a /h/-V-/d/ context, giving the words: *heed, hid, head, had, hard, hod, hoard, who'd, hood, hud, heard, hayed, hide, how'd, hoed, haired, hoyed*. These words were embedded in the carrier sentence "*Say __ again*". Recordings were made under identical conditions and using the same equipment as the consonant recordings. Again, each word was manually checked for clarity and the clearest one was chosen for the stimuli.

The selected recordings were used to create stimuli for three experimental conditions: quiet, natural vowels in noise, and duration equated vowels in noise. The latter condition was included to test the use of duration as a cue in vowel identification; the Arabic vowel inventory includes short-long pairs, and so it is possible that Arabic learners are able to make use of duration as an L1 cue when identifying English vowels. Duration equated vowels were created using PSOLA implemented in Praat (Boersma & Weenink, 2005). The duration of the /h/ closure, the duration for the vowel, and the duration of /d/ closure were averaged across all vowels for the talker,

and then these values were used for all words. To create the stimuli for the different noise conditions, recordings were equated for amplitude and then speech-shaped noise was added to the natural and duration-equated recordings to create three SNRs (0, -5, -10 dB).

3.3.3 Procedure

All perception experiments were carried out in sound-attenuated audio-booths at UCL Language Sciences, Chandler House. Stimuli were presented over Sennheiser HD 555 headphones and both stimuli presentation and response collection was controlled using PRAAT (Boersma & Weenink, 2005).

3.3.3.1 Consonant perception

Listeners completed two tasks: consonant identification in quiet and consonant identification in noise. L2 participants listened to recordings of the English speaker for the consonants in the two vocalic contexts: / ɑCɑ/, and /iCi/ in the carrier sentence: *say__ again* (e.g., "*Say /aka/ again*", "*Say /aʤa/ again"),* and were asked to give a closed-set identification response with all 23 words as response options. Only L2 listeners were tested in the quiet condition.

3.3.3.2 English Consonant identification in quiet

Participants were presented with an on-screen display showing all 23 consonants with example words, such as "*B as in Bear*", "*SH as in Sharp*". Words were selected to be high frequency and pilot testing confirmed that they were familiar to all participants regardless of L2 proficiency. Before completing the experiment, participants were familiarized with the task and materials. They were given instructions on how the task would proceed, and in particular were familiarised with words where the acoustic-orthographic correspondence is not transparent, (e.g., 'th' can be produced as /ð/ as in *faTHer*, or as /θ/ as in *THeatre*) (see Fig. 3.1)

Figure 3.1: Screenshot of the consonant identification, participants were presented with this screen for identification in quiet and in noise

To control for any training or order effects, participants were allocated randomly to two testing blocks; half of the participants started with the /iCi/ and half with the /ɑCɑ/ context. Participants identified three repetitions of each consonant in each context, giving a total of 138 responses (23 consonants x 3 repetitions x 2 vocalic contexts, giving 69 stimuli for each vocalic context), with the order of presentation within each block randomized. The test was self-paced with a break mid-way through the tasks (i.e., after 69 stimuli).

3.3.3.3 English Consonant identification in noise

This task was completed by both non-native (L2) and native L1 (SSBE control) listeners. Listeners identified two repetitions of each consonant in two vocalic contexts (/ɑCɑ / and /iCi/) and at three different SNRs: 0 dB, -5dB and -10dB. This gave a total of 46 stimuli for each vocalic context, and a total of 92 stimuli per noise condition. The experiment was blocked by noise level and the order of presentation of the blocks randomized to control for any learning effects. Additionally, the order of

presentation of the stimuli was randomized within each block. The test was self-paced with a break mid-way through (i.e., after 138 stimuli).

Vowel perception

Listeners completed three tasks: natural vowel identification in quiet, natural vowel identification in noise and duration equated vowel identification in noise. As for consonant perception, only non-native (L2) listeners were tested in the quiet condition.

Vowel identification in quiet. Participants listened to recordings of the vowels in /hVd/ words in the carrier sentence "*Say_ again"*, and were asked to give a closed-set identification response from the 17 test words. The stimuli were presented with a screen layout showing the 17 vowels as their /hVd/ words along with a rhyming word, (e.g., *heed* as in *seed*, *hud* as in *cut*). As for the consonantal stimuli, these words were selected to be high frequency and pilot testing confirmed that they were familiar to all participants regardless of L2 proficiency. Listeners identified three repetitions of each vowel in a randomized order, giving a total of 51 trials. The test was self-paced with no break (see Fig. 3.2).

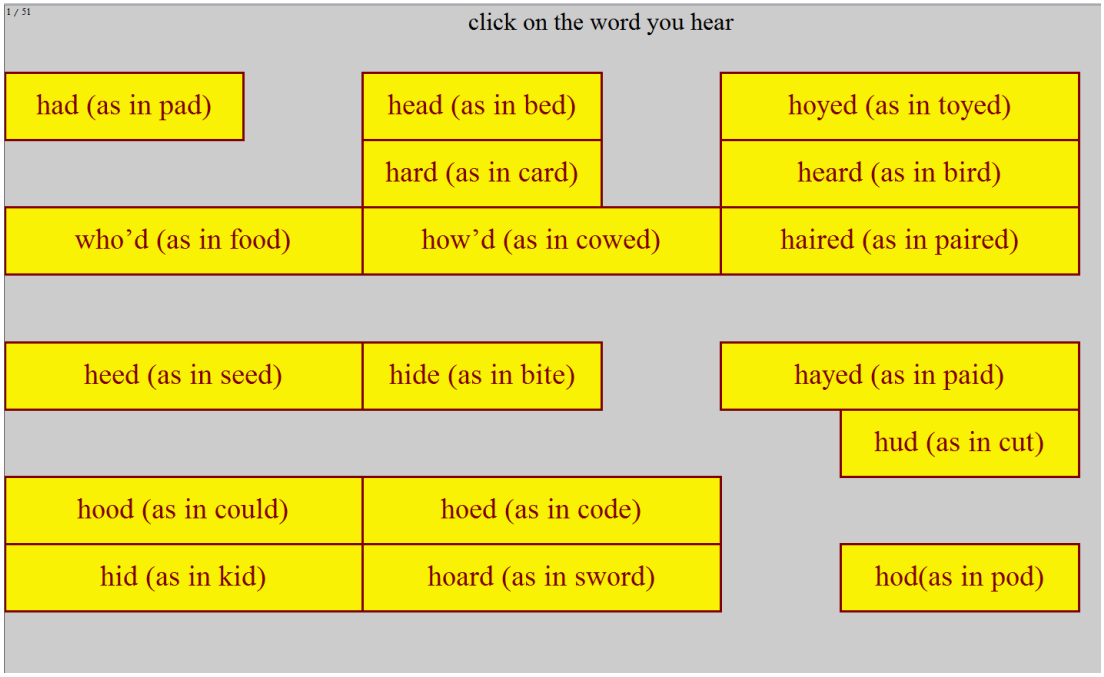


Figure 3.2: Screenshot of the vowel identification task, participants were presented with this screen in vowel identification in quiet, noise and duration equated tasks.

Vowel identification in noise. This task was completed by both non-native (L2) and native L1 (SSBE control) listeners. Listeners identified two repetitions of each vowel at the three different SNRs: 0 dB, -5 dB and -10 dB (the repetitions were reduced to 2 given that each vowel is repeated at each SNR level). This gave a total of 102 stimuli (17 vowels x 3 SNR levels x 2 repetitions, giving 34 stimuli per noise level). The experiment was blocked by noise level and the order of presentation of the blocks randomized to control for any learning/order effects. Additionally, the order of presentation of the stimuli was randomized within each block. Responses were collected using the same procedure used in the vowel identification in quiet test. The test was self-paced with a break mid-way through (i.e., after 51 stimuli).

Identifying duration equated vowels in noise. This task was completed by both non-native (L2) and native L1 (SSBE control) listeners. Listeners identified two repetitions of each vowel at the three different SNRs: 0 dB, -5dB and -1dB. This gave a total of 102 stimuli (34 stimuli per noise level). The experiment was blocked by noise level and the order of presentation of the blocks randomized to control for any learning/order effects. Additionally, the order of presentation of the stimuli was randomized within each block. Responses were collected using the same procedure used in the 'vowel identification in quiet' test. The test was self-paced with a break mid-way through (i.e., after 51 stimuli).

Vowel Production

Recordings. After completing the perception tasks, the non-native (Saudi) participants recorded the same 17 vowels they were asked to identify in the vowel perception task. Participants recorded three repetitions of each of the /hVd /words in the carrier sentence *Say __ again*. Stimuli were presented via PowerPoint, one word per slide. To obtain a sample of their connected speech, participants also recorded the phonetically balanced paragraph "The north wind and the sun" (IPA Handbook, 1999). Participants were instructed to read the passage twice before recording, in order to minimise mistakes or disfluencies during recording. They were also instructed to read this at a conversational speed. The paragraph was also presented via PowerPoint. All recordings were made using a C1U USB microphone in a sound-attenuated room at a

sampling rate of 44100 Hz (16-bit) samples/s and were later down sampled to 22050 Hz.

Vowel intelligibility and accent ratings. Native SSBE listeners identified vowels and rated samples of the Arabic speakers' speech. All participants were tested in a sound-attenuated room using PRAAT (Boersma and Weenink, 2005). Stimuli were presented using Sennheiser HD 555 headphones at a user-controlled comfortable level.

Vowel intelligibility. Vowel repetitions were checked for clarity, and for each speaker the best repetition (i.e., clear voice quality, no hesitation) was chosen as the stimulus for the intelligibility task. This gave a total of 442 stimuli: 17 vowels per non-native speaker. Nine native SSBE listeners identified Arabic speakers' vowels. They were presented with the same screen layout that was used in perceptual task on which the 17 vowels were represented in /h/-V-/d/ words with rhyming words (e.g., *heed* as in *seed)*. The order of the stimuli and the talker was randomised, and the identification task was self-paced with participant-controlled breaks after 50 stimuli.

Accent ratings. Nine native SSBE listeners rated an extract of the Arabic speakers' recordings of "The North Wind and the Sun". The same extract was taken from each recording: "*Then the North Wind blew as hard as he could, but the more he blew, the more closely did the traveller fold his cloak around him; and at last the North Wind gave up the attempt*". This extract was selected because it contains a range of vowels and in particular, consonant clusters which are not phonotactically permissible in Arabic. The rating sessions were self-paced and listeners could listen to each extract twice; the order of the extracts was randomised. Listeners gave their ratings on a 7-point Likert scale where 1 was judged to be very native-like, and 7 very non-native.

3.4 Results

Results were analysed for each task separately with L2 learners split into two groups: high proficiency (HP) and low proficiency (LP). Participants were divided to the HP or LP group based on their score in the Oxford English Grammar Test (Allan, 1992). Participants were assigned to the high proficiency group if their score was

higher than or equal to the median score and the low proficiency group if their score was lower than the median (range of scores 17- 47 out of 50, median 29.5).

3.4.1 Consonant perception

English Consonant identification in quiet

Figure 3.3 displays the accuracy for consonant identification in quiet for HP and LP groups. As might be expected, the LP group appeared to perform more poorly than the HP group. This observation was tested using an independent samples t-test. The performance of the two groups was significantly different, [$t=3.6$, $p<.05$, $df=24$], confirming that proficiency level was a significant factor determining L2 listeners' ability to perceive L2 phonemes: HP listeners were more accurate in their identification performance than LP listeners.

A series of analyses investigated whether perceptual confusions were affected by proficiency. Table 3.1 displays the confusion matrix for the LP group. Participants were very accurate with some phonemes but performed more poorly with affricates /tʃ, dʒ/ fricatives /ʃ, ʒ/ the dental fricatives /θ-ð/ and the velar nasal /ŋ/. Table 3.2 displays the confusion matrix for the HP group. Likewise, this group were very accurate with some phonemes but performed more poorly with affricates /tʃ- dʒ / and fricatives /ʃ- ʒ/, and the velar nasal /ŋ/. Performance on the dental fricatives /θ- ð/ was also slightly lower than for other phonemes.

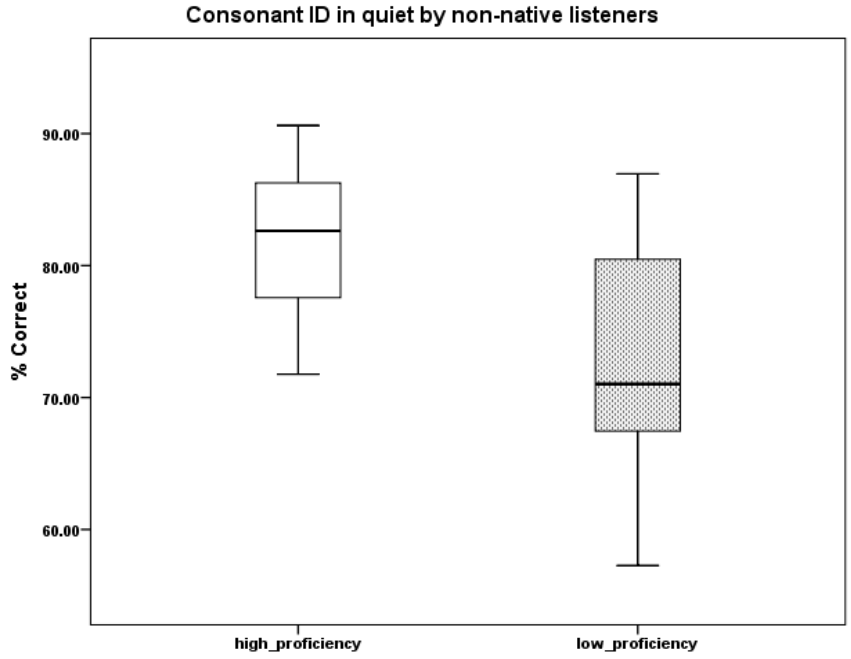


Figure 3.3: Boxplot showing the consonant identification accuracy (percentage correct) in quiet averaged across vocalic contexts and split into high proficiency and low proficiency groups.

Separate hierarchical cluster analyses for the HP and LP groups were used to analyse the confusion patterns. The resulting analysis is shown graphically in Figures 3.4 & 3.5. The strength of clustering is indicated by the level of similarity at which elements join the cluster. Thus, phonemes that join at a similar level are considered to have a similar level of confusability. Clusters that are formed lower down the scale are more confusable than those formed higher up the scale.

Figure 3.4 demonstrates that there were four distinct clusters for the LP group; one containing the affricates, postalveolar fricatives and closest voiced stop /g/, another containing the dental fricatives and voiceless labio-dental fricative, another made up of the alveolar and velar nasals, and lastly, a cluster made up of the bilabial plosives. Within each of these clusters, certain groups of consonants were highly confusable; the affricate /dʒ/, and the fricative /ʒ/ were the most confusable and joined to form the first cluster. The alveolar nasal /n/, and the velar nasal /ŋ/ form the second

cluster, the voiceless affricate /tʃ/ and the postalveolar fricative /ʃ/ were also highly confusable. The dental fricatives /θ/ and /ð/ and bilabial plosives /b/ and /p/ were less confusable.

The cluster diagram for the HP group displays some differences from the LP group analysis (see Fig. 3.5). There are two clusters, one containing the voiced affricate /dʒ/ and corresponding voiced fricative /ʒ/, and the other one containing the alveolar nasal /n/ and velar nasal /ŋ/. The analysis indicates that the voiced affricate /dʒ/ and the voiced fricative /ʒ/ were the most confusable phonemes and joined to form the first cluster, followed by the alveolar nasal /n/ and the velar nasal /ŋ/. It is worth noting that these two clusters were problematic contrasts for both proficiency groups.

	response																								
		b	tʃ	d	dʒ	f	g	h	k	l	m	n	ŋ	p	r	s	ʃ	t	θ	ð	v	w	z	ʒ	Total
	b	68	0	1	0	0	0	0	0	0	0	0	0	27	0	0	0	0	0	0	4	0	0	0	100
	tʃ	0	33	0	9	0	17	0	0	0	0	0	5	0	1	0	29	0	0	0	0	0	0	5	100
	d	0	0	87	0	0	8	0	0	0	0	0	3	0	0	0	0	1	0	1	0	0	0	0	100
	dʒ	1	0	0	31	0	36	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	23	100
	f	0	0	1	0	82	0	0	0	0	0	0	0	0	0	0	0	0	4	5	8	0	0	0	100
	g	0	0	3	0	0	86	0	1	0	0	0	8	0	0	0	0	0	0	0	3	0	0	0	100
	h	0	0	0	1	1	0	87	0	0	0	0	0	0	0	1	4	0	0	3	1	1	0	0	100
	k	0	4	0	0	0	3	0	91	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	100
	l	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
	m	0	0	0	3	0	1	1	1	0	79	10	0	0	0	0	0	1	1	0	0	0	1	0	100
	n	0	0	0	0	0	0	0	0	0	0	79	21	0	0	0	0	0	0	0	0	0	0	0	100
	ŋ	0	0	1	1	0	0	0	0	0	3	47	47	0	0	0	0	0	0	0	0	0	0	0	100
	p	17	0	0	0	0	0	0	0	0	0	0	0	74	3	0	0	0	0	1	4	0	0	1	100
	r	0	0	0	0	0	0	0	0	0	0	0	0	0	99	0	0	0	0	0	0	1	0	0	100
	s	0	24	0	0	0	0	0	0	0	0	0	1	0	0	71	3	0	0	0	0	0	1	0	100
	ʃ	0	22	0	1	0	0	1	0	0	0	0	0	0	0	0	74	0	0	0	0	0	0	1	100
	t	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	92	8	0	0	0	0	0	100
	θ	0	0	0	0	27	0	0	0	0	0	0	0	0	0	0	3	1	54	14	1	0	0	0	100
	ð	0	0	0	0	4	0	0	0	0	0	1	0	0	0	0	1	0	19	64	8	0	1	1	100
	v	1	0	0	0	8	0	0	0	0	0	0	0	0	0	0	1	0	0	3	87	0	0	0	100
	w	0	0	0	1	0	0	0	0	0	0	0	0	1	13	0	0	0	0	0	0	85	0	0	100
	z	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	1	0	94	3	100
stimulus	ʒ	0	0	0	21	0	42	0	1	0	0	0	8	0	0	0	0	0	0	0	0	0	0	28	100

Table 3.1: Consonant Confusion matrix for the low proficiency group (LP); the stimuli are in rows, and the responses (Percentage correct) in columns. Responses are averaged over both vocalic contexts

		response																						
		b	**tʃ**	**d**	**dʒ**	**f**	**g**	**h**	**k**	**l**	**m**	**n**	**ŋ**	**p**	**r**	**s**	**ʃ**	**t**	**θ**	**ð**	**v**	**w**	**z**	**ʒ**	**Total**
	b	91	0	0	1	0	0	0	0	0	0	0	0	6	0	0	1	0	0	0	0	0	0	0	100
	tʃ	0	76	0	1	0	4	3	0	0	0	0	0	0	0	0	13	0	0	0	0	0	0	4	100
	d	0	0	92	0	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
	dʒ	0	1	0	59	0	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17	100
	f	0	0	0	0	90	0	0	0	0	0	0	0	0	0	0	0	0	3	5	3	0	0	0	100
	g	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
	h	0	0	0	1	0	0	99	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
	k	0	0	1	0	0	0	0	97	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	100
	l	0	0	0	0	1	0	0	0	99	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
	m	0	0	0	0	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	100
	n	0	0	0	0	0	0	0	0	0	0	94	6	0	0	0	0	0	0	0	0	0	0	0	100
	ŋ	0	0	0	0	0	4	0	0	1	0	35	60	0	0	0	0	0	0	0	0	0	0	0	100
	p	8	0	0	0	0	0	0	1	0	0	0	0	86	0	0	1	0	0	0	0	0	1	3	100
	r	0	0	1	0	0	0	0	0	0	0	0	0	0	96	1	0	0	0	0	0	0	1	0	100
	s	0	8	0	0	0	0	0	0	0	0	0	0	0	3	90	0	0	0	0	0	0	0	0	100
	ʃ	0	9	0	0	0	0	0	0	0	0	0	0	0	0	1	87	0	0	1	0	0	0	1	100
	t	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	95	4	0	0	1	0	0	100
	θ	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	74	6	0	0	0	0	100
	ð	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	9	79	9	1	0	0	100
	v	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	6	91	0	0	0	100
	w	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	99	0	0	100
	z	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	97	1	100
Stimulus	**ʒ**	0	0	0	46	0	22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	33	100

Table 3.2: Consonant Confusion matrix for the high proficiency group (HP); the stimuli are in rows, and the responses (Percentage correct) in columns. Responses are averaged over both vocalic contexts

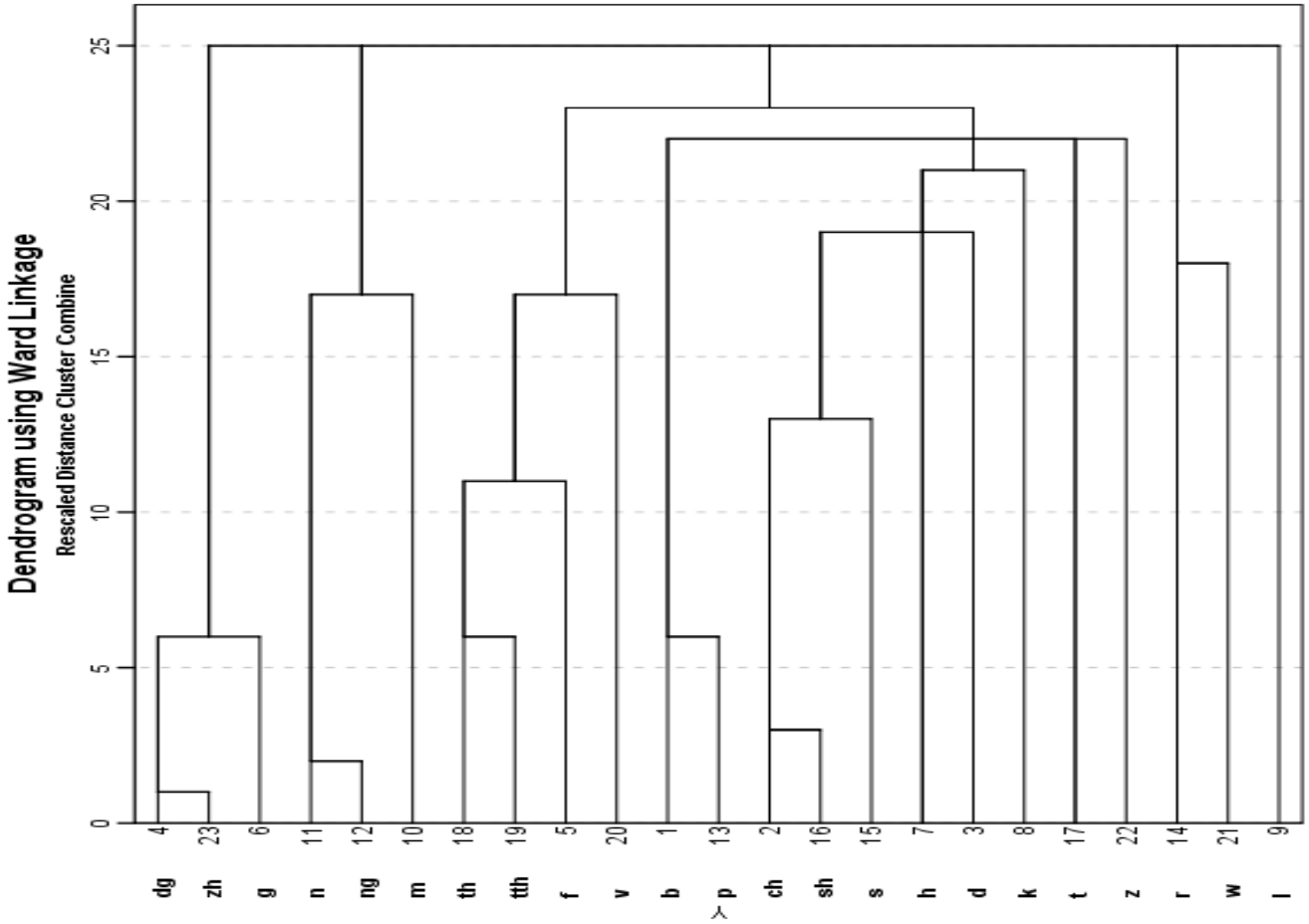


Figure 3.4: Clustering solution for the nearest neighbours in the confusion matrix for the LP group. The y-axis shows the distance between clusters, and the x-axis shows the consonants and how close/far they are confused, (th=θ, tth=ð)

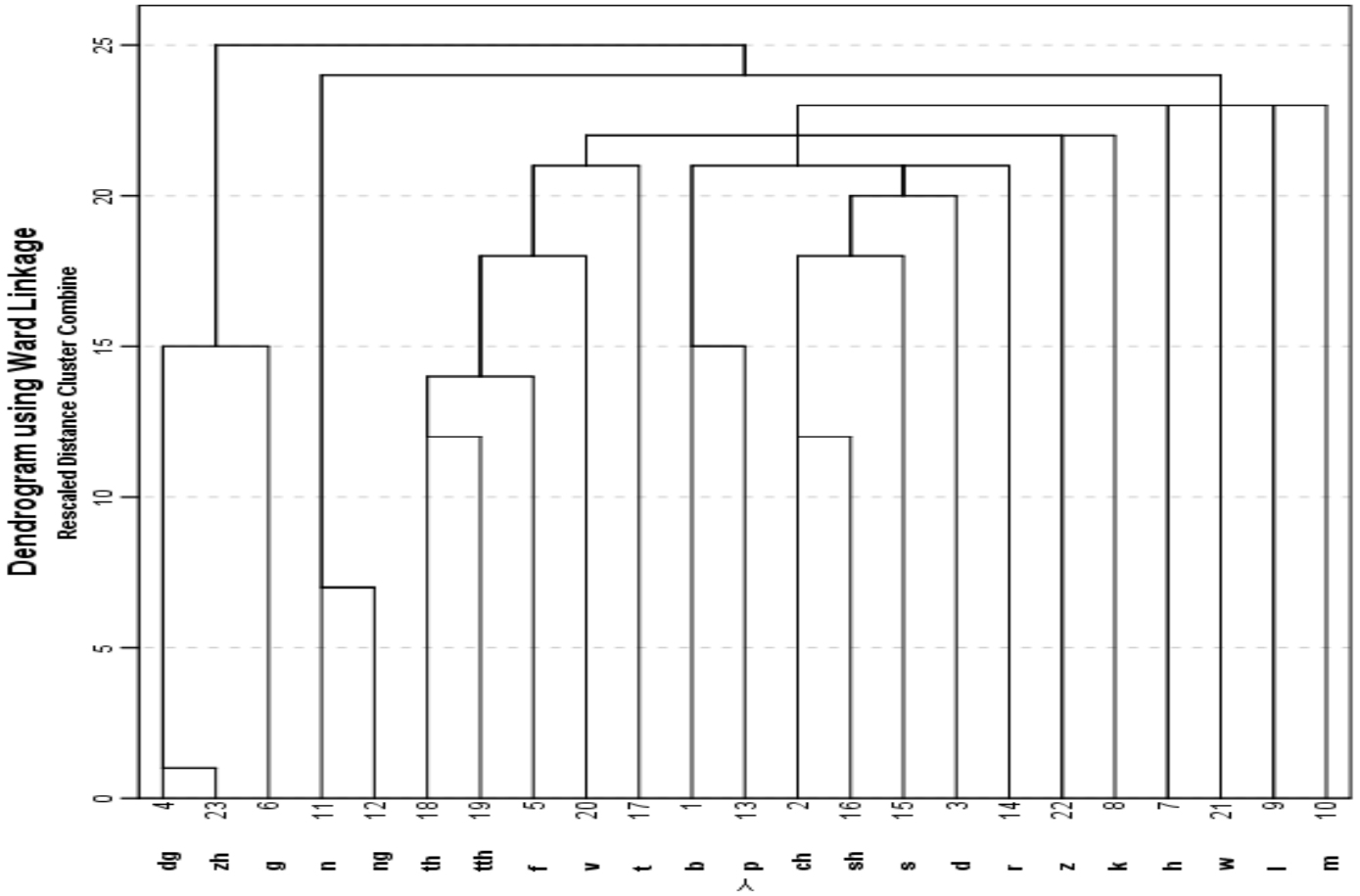


Figure 3.5: Clusters of the distance between the nearest neighbours in the confusion matrix for the HP group, y-axis shows the distance between clusters, and x-axis shows the consonants and how close/far they are confused, (th=θ, tth=ð).

English consonants in noise

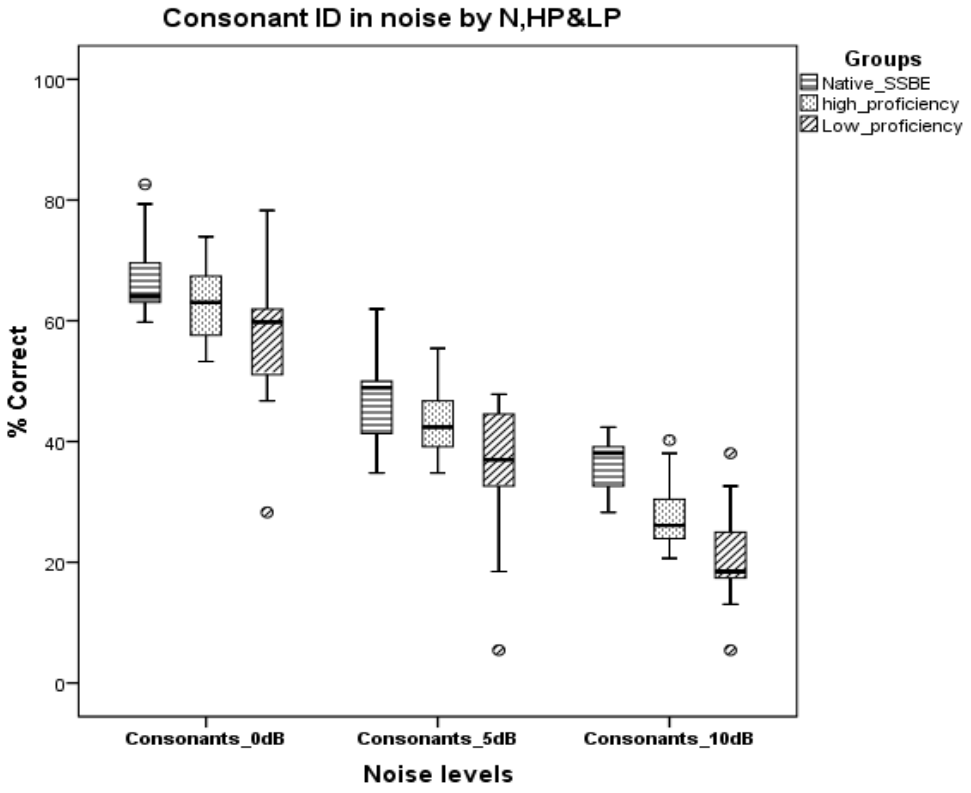


Figure 3.6: Boxplot to show consonant identification (percentage correct) in three different noise levels (0, -5, -10 dB) for three groups, natives (SSBE), high, and low proficiency (Arabic) listeners, averaged across vocalic conditions.

Figure 3.6 displays the English consonant identification accuracy in noise for each group: natives (NT), high proficiency (HP), and low proficiency (LP) non-natives. As expected, all listeners performed more poorly at higher noise levels. Performance appeared to be affected by proficiency and listener groups appeared to be equally affected by noise: NT listeners performed best, followed by HP and then LP listeners, and performance did not appear to drop more for non-native than for native listeners in the higher noise conditions.

These observations were tested using a repeated measures ANOVA with noise level (0dB, -5dB, -10dB) coded as a within-subjects factor, and group (NT vs. HP vs. LP) as a between-subjects factor. The main effect of noise was significant [$F(2,64)=258.98$, $p<.001$], confirming that overall performance differed according to noise level performance decreased as the noise level increased (0dB: 61%, -5dB: 41%, -10dB: 28%). There was a significant main effect of group, [$F(1,32)=7.1$, $p<.05$]; overall performance accuracy for the NT listeners was higher (49.9%) than for the HP group (44.8%), and the LP group (36.7%). As expected, the LP group performed more poorly than the HP group, and the HP group performed worse than the NT listeners in noise. However there was no interaction between group and noise, indicating that all listeners were affected by the noise.

3.4.2 Vowel perception

English vowel identification in quiet

Figure 3.7 displays the accuracy for English vowel identification in quiet for HP and LP listeners. As expected the HP group performed better than the LP group. An independent samples t-test revealed that there was a significant difference between the HP and LP group [$t=2.72$, $p<.05$, $df=24$], confirming that HP learners identified English vowels more accurately than the LP group.

A series of analyses investigated whether perceptual confusions were affected by proficiency. Table 3.3 displays the confusion matrix for the LP group. Participants were accurate with some phonemes (e.g., /iː/ *heed,* /æ/ *had,* /ɑː/ hard) but performed particularly poorly with the following vowels; the front-mid vowel /ɪ/ (*hid*), the high-back vowel /uː/ (*who'd*), the mid-back vowel /ʊ/ (*hood*) , the mid closing diphthong /əʊ/ (*hoed*), the open-back vowel /ɒ/ (*hod)*, the low central vowel /ʌ/ (*hud*), the mid-central vowel /ɜː/ (*heard*) and the central diphthongs /ɛə/ *(haired)* and /əʊ/ (*hoed*). Table 3.4 displays the confusion matrix for the HP group. These participants had fewer difficulties overall, but still found some of the same vowels problematic; the front-mid vowel (/ɪ/ *hid*), the open-back vowel (/ɒ/ *hod*) and the low central vowel (/ʌ/ *hud*), and the mid-central vowel (/ɜː/ *heard*) and the central diphthongs (/ɛə/ *haired and /əʊ/ hoed*).

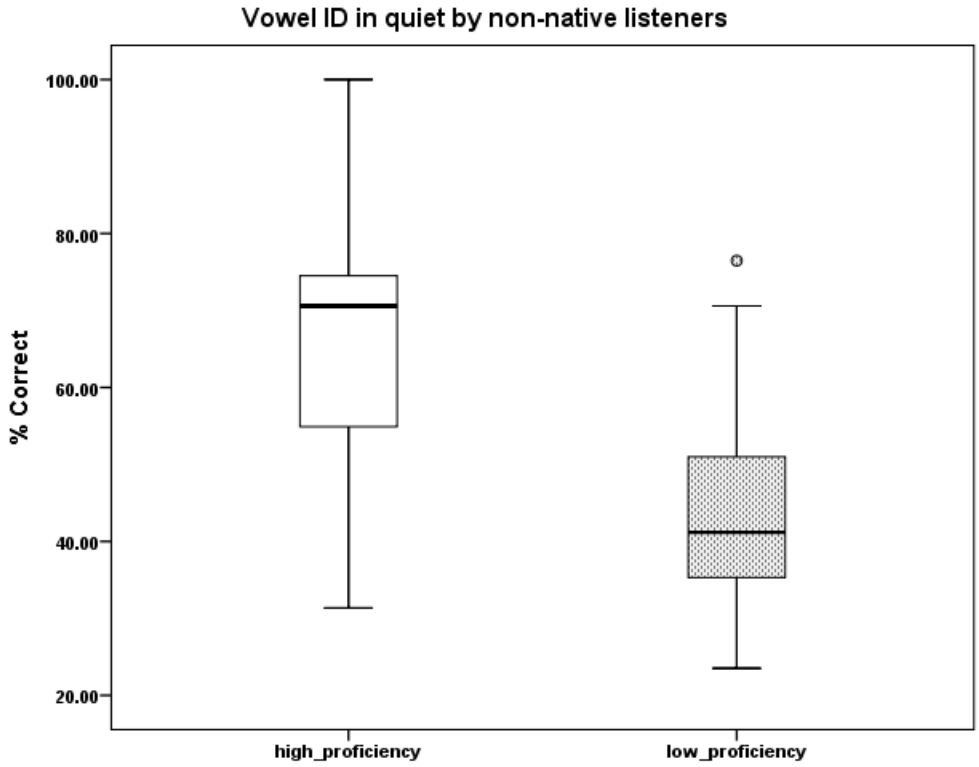


Figure 3.7: Boxplot to show the vowel identification accuracy (percentage correct) for high and low proficiency groups. High proficiency learners performed better overall than did low proficiency learners.

Separate hierarchical cluster analyses for the HP and LP groups were used to analyse the confusion patterns. The resulting analysis is presented in Figure 3.8. For the LP group there were three distinct clusters; high front vowels, high-back vowels, and central and low back vowels. Within these clusters, certain pairs of vowels were highly confusable; the front-mid vowel /ɪ/ (*hid*) and the open-mid vowel /ɛ/ (*head*) were the most confusable contrasts and joined to form the first cluster, followed by the high-back vowel /uː/ (*who'd*) and mid-back vowel /ʊ/ (*hood*), and the diphthong /əʊ/ (*hoed*). The open-back vowel /ɒ/ *(hod*) and the low central vowel /ʌ/ (*hud*) were also highly confusable, as were the mid-central vowel /ɜː/ (*heard*) and central diphthong /ɛə/ *(haired*).

The resulting analysis for the HP group in Figure 3.9 shows three distinct clusters that contained vowel confusions; high back vowels /u/ (*hood)* and /əʊ/ *(hoed*), and low-back vowel /ɒ/ (*hod*) and the central vowel /ʌ/ (*hud*). Within these clusters, the open-back vowel /ɒ/ (*hod*) the low central vowel /ʌ/ (hud), and the diphthong /əʊ/ (*hoed*) were the most confusable contrasts and joined to form the first cluster, followed by the front-mid vowel /ɪ/ (*hid*) and the open-mid vowel /ɛ/ (*head*), the mid-central vowel (/ɜː/ *heard*), and the diphthong /ɛə/ (*haired*). The last cluster contained the least confusable vowel contrasts the diphthongs /əʊ/ (*hoed*), the central vowel /ʌ/ (*hud*), and the back vowel /ʊ/ (*hood*).

Although the HP group had fewer difficulties overall, they shared some of the same vowel confusions with the LP group; high front vowels /ɪ/ (*hid)* and /ɛ/ (*head*), central vowels /ɜː/ (heard) and /ɛə/ (haired) and low back and central vowels /ɒ/ (*hod),* /ʌ/ *(hud), and* /əʊ/ *(hoed*). The high back vowel, /ʊ/ (*hood*), also seemed to present some difficulties. This vowel was confused with /ɒ/ *(hod*), a low front vowel. This was surprising as these vowels are in different parts of the vowel space. It is possible, however, that this is a result of orthographic rather than phonetic interference. That is, when L2 learners heard the word "*hood*" they responded by clicking on the word "*hod*" as they associated the double "*oo"* with the long /u/ as in (*food*).

		response																	
		iː	ɪ	e	æ	ɑː	ɒ	ɔː	uː	ʊ	ʌ	ɜː	eɪ	aɪ	aʊ	əʊ	ɛə	ɔɪ	Total
	iː	74	8	3	0	0	0	0	0	0	0	0	3	10	0	3	0	0	100
	ɪ	3	8	72	5	0	0	0	0	0	0	3	3	5	0	0	0	3	100
	e	5	10	69	10	0	0	0	0	0	0	0	3	3	0	0	0	0	100
	æ	0	0	0	79	3	0	0	0	0	13	0	0	3	3	0	0	0	100
	ɑː	0	0	0	5	85	3	3	0	0	3	3	0	0	0	0	0	0	100
	ɒ	0	0	3	0	38	3	5	0	5	18	3	0	0	3	13	10	0	100
	ɔː	0	0	0	0	3	5	62	3	5	3	3	0	0	10	0	0	8	100
	uː	0	0	0	0	0	5	0	36	54	0	0	0	0	5	0	0	0	100
	ʊ	0	0	0	0	0	10	0	23	51	8	0	0	0	0	8	0	0	100
	ʌ	0	0	3	5	21	15	0	0	0	31	21	3	0	0	3	0	0	100
	ɜː	0	0	5	0	23	0	3	0	0	0	44	3	0	0	0	23	0	100
	eɪ	0	0	18	0	8	0	0	0	0	0	0	59	5	0	0	3	8	100
	aɪ	0	26	3	0	0	0	0	0	0	0	0	18	46	0	0	3	5	100
	aʊ	0	0	0	0	0	8	3	10	13	0	0	3	0	59	5	0	0	100
	əʊ	0	0	0	0	0	8	0	5	36	8	0	0	0	26	18	0	0	100
	ɛə	0	8	21	21	3	0	0	0	0	0	28	3	0	0	0	18	0	100
stimulus	ɔɪ	0	0	0	0	0	0	3	0	0	0	3	5	0	5	26	3	56	100

Table 3.3: Vowel confusion matrix for the LP group listeners. The stimuli are in rows, and the responses (percentage correct) in columns

		Response																	
		iː	ɪ	e	æ	ɑː	ɒ	ɔː	uː	ʊ	ʌ	ɜː	eɪ	aɪ	aʊ	əʊ	ɛə	ɔɪ	Total
	iː	86	0	6	0	0	0	0	0	0	0	6	3	0	0	0	0	0	100
	ɪ	0	44	47	3	0	3	0	0	0	0	0	0	3	0	0	0	0	100
	e	0	11	83	3	0	0	0	0	0	0	0	0	3	0	0	0	0	100
	æ	0	0	0	86	0	6	0	0	0	8	0	0	0	0	0	0	0	100
	ɑː	0	0	0	0	83	3	0	0	3	0	6	0	0	0	0	6	0	100
	ɒ	0	0	0	0	8	31	6	0	14	11	0	0	0	8	22	0	0	100
	ɔː	0	0	0	3	0	0	72	3	3	0	0	0	0	17	0	0	3	100
	uː	0	0	0	0	0	3	0	61	19	0	0	0	0	8	3	0	6	100
	ʊ	0	3	0	3	0	14	3	6	69	0	0	0	0	0	3	0	0	100
	ʌ	0	3	0	8	3	14	0	0	8	44	11	3	0	3	3	0	0	100
	ɜː	0	0	3	0	17	0	3	0	0	0	78	0	0	0	0	0	0	100
	eɪ	0	0	3	0	0	0	0	0	0	0	6	86	3	0	0	3	0	100
	aɪ	0	8	0	0	0	0	0	0	0	0	0	0	92	0	0	0	0	100
	aʊ	0	0	0	0	0	3	3	3	3	0	0	0	0	78	8	0	3	100
	əʊ	0	0	0	0	0	0	0	6	19	0	0	3	0	17	53	0	3	100
	ɛə	0	0	17	6	3	0	3	0	0	0	36	3	3	0	0	31	0	100
Stimulus	ɔɪ	0	0	0	0	0	0	3	0	0	0	0	6	0	3	0	0	89	100

Table 3.4: Vowel confusion matrix for the HP group listeners. The stimuli are in rows, and the responses (percentage correct) in columns.

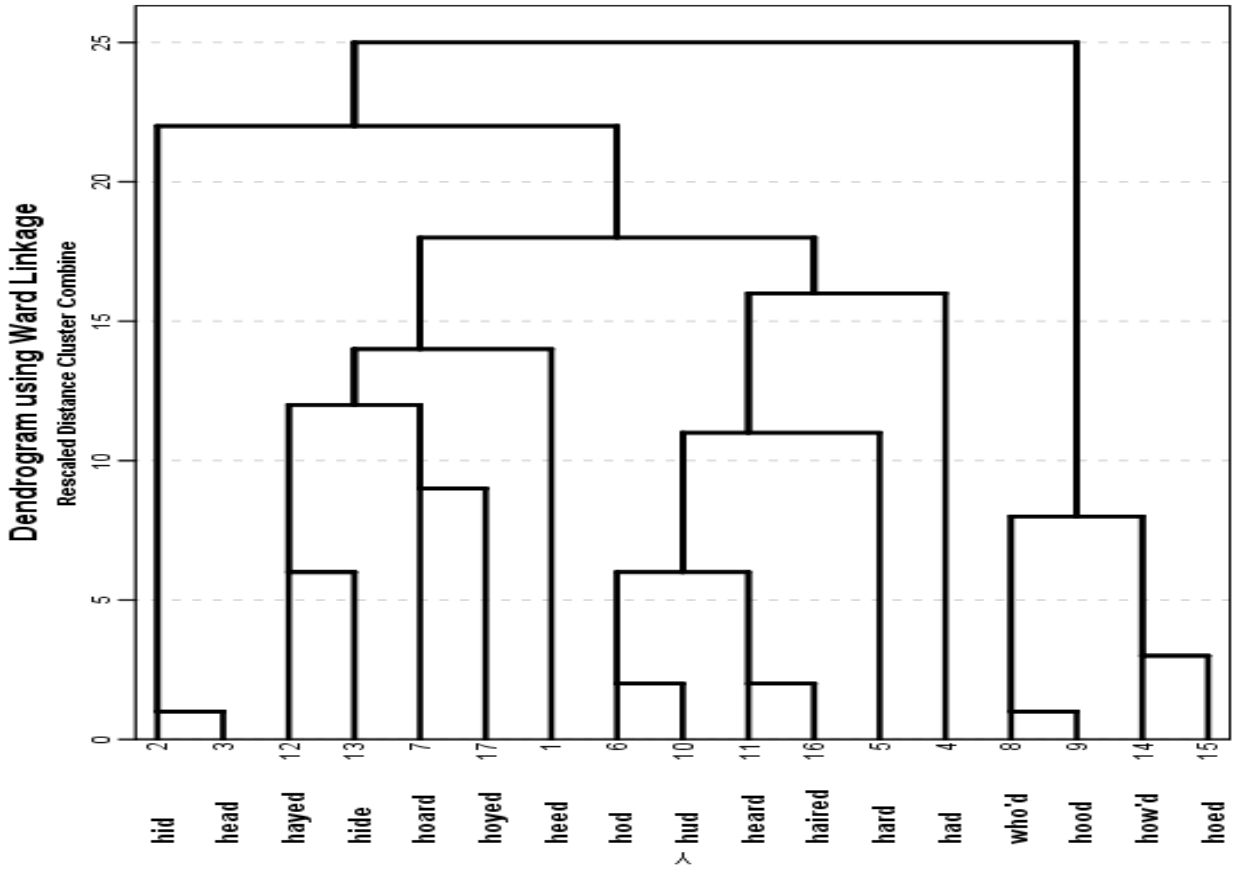


Figure 3.8: Clustering solution showing the distance between the nearest neighbours in the confusion matrix for the LP group; the y-axis shows the distance between clusters, and the x-axis shows the vowel categories.

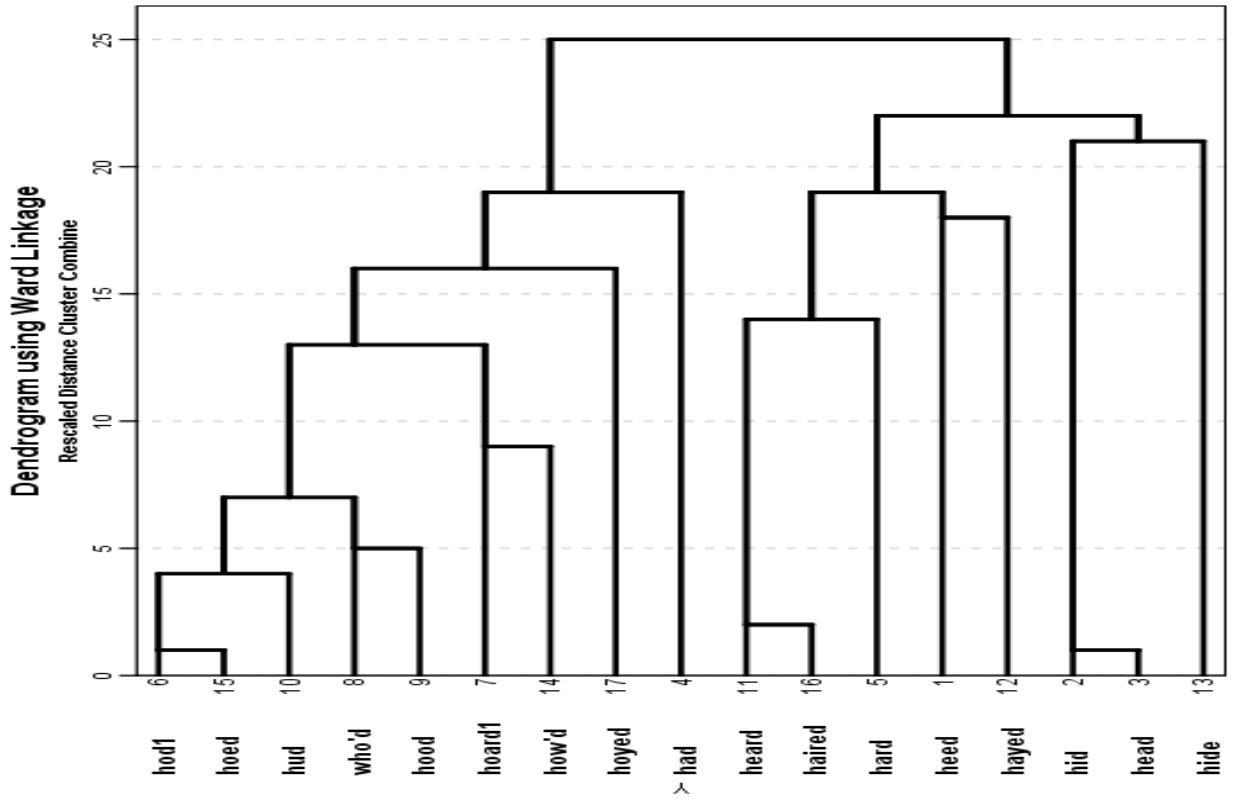


Figure 3.9: Clustering solution showing the distance between the nearest neighbours in the confusion matrix for the HP group; the y-axis shows the distance between clusters, and the x-axis shows the vowel categories

Vowel identification in noise natural vs. duration equated vowels

Figure 3.10 displays the accuracy performance for the three groups; native SSBE (NT), HP, and LP groups. A repeated-measures ANOVA examined the effect of group, duration equated vs. natural vowels, and SNR on performance accuracy. We predicted that overall non-native listeners in general would perform worse than natives, and that the LP would perform worse than the HP group in noise. Since duration is contrastive in Arabic, we hypothesized that Arabic listeners, especially the LP group, might rely on duration more when identifying vowels that are not present in their L1 (e.g., head) when identifying vowels than would native (SSBE) listeners, who are thought to rely more on spectral rather than duration information (see e.g., Escudero & Boersma, 2004; Escudero et al. 2009).

To measure the possible differences between groups, a repeated measures ANOVA was run with duration (natural, duration equated), and noise (0dB, -5dB, -10dB) as within-subjects-factors, and group (NT vs. HP vs. LP) as a between-subjects factor. The main effect of duration was significant [$F(1,32)= 17.51$, $p<.001$]; overall identification of natural vowels averaged across all listeners was higher for natural vowels (52%) than for the duration equated vowels (44%), indicating that participants found the natural vowels easier to identify. The main effect of noise was significant [$F(2,64)=21.7$, $p<.001$], overall performance differed according to noise level (0dB:59%, -5dB:5%, -10dB:35%), demonstrating that performance dropped as the noise level increased. As expected, the main effect of group was significant [$F(1,32)=31.78$, $p<.001$]; overall performance for the NT listeners was higher (7.69%) than for the HP (46%) and LP groups (29%). There was no interaction between duration and groups ($p > .05$) indicating that LP Arabic learners did not rely more on duration when identifying vowels. However there was a significant two-way interaction between noise and groups [$F(4,64)=13.62$, $p<.001$]. Inspection of the data revealed that this was because NT listeners were more affected by the higher noise levels more than were the non-natives, who performed more poorly at the easier noise levels.

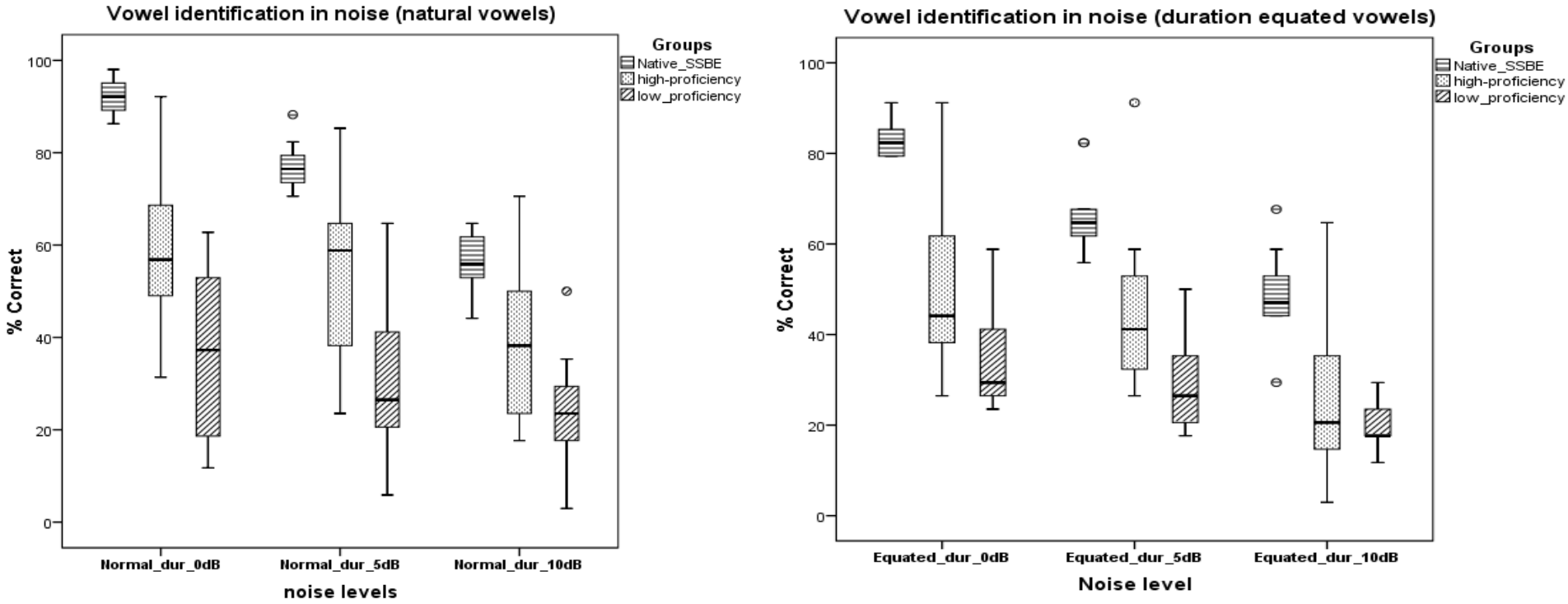


Figure 3.10: Boxplots showing the overall vowel identification scores (percentage correct) for the three groups (N, HP, and LP) in natural vowels, and in the duration equated condition at the three noise levels (0, -5, and -10 dB)

3.4.3 Production tasks

Vowel intelligibility

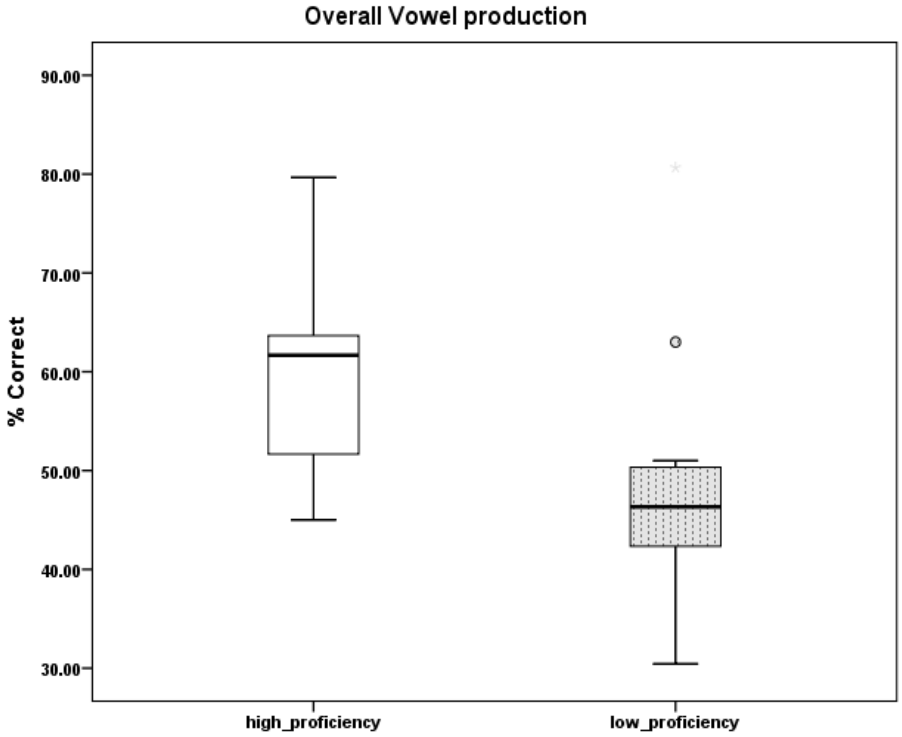


Figure 3.11: Boxplot showing overall vowel identification (percentage correct) of L2 speakers' productions identified by SSBE listeners

Figure 3.11 suggests that the HP speakers were more intelligible than the LP speakers. This was confirmed by an independent samples t-test which confirmed that performance accuracy was significantly higher for the HP than the LP speakers [t=2.94, p<.05, df=24].

To investigate whether Arabic speakers found particular vowel contrasts difficult to produce, the data were submitted to confusion matrices. Table 3.5 shows the confusion matrix for the vowels produced by the LP group, and identified by the NT listeners. The NT listeners frequently confused /ɪ/ (*hid)* and /ɛ/ (*head),* /ɑː/ *(hard),* /ɔː/ (*hoard*) and /ɜː/ (*heard)*, /uː/ *(who'd)* and /ʊ/ (*hood),* /eɪ/ *(hayed)* and /aɪ/ (*hide).*

The diphthong /əʊ/ *(hoed)* was often misidentified as /ʊ/ (*hood),* and the diphthong /ɛə/ (*haired)* as /ɜː/ (*heard).* Table 3.6 displays the confusion matrix for the vowels produced by the HP group. Similarly, the NT listeners confused the vowels that were produced by the HP speakers; /ɪ/ *(hid) and* /ɛ/ *(head)* and /uː/ *(who'd)* and /ʊ/ (*hood).* The low back vowel /ɒ/ (*hod)* was misidentified as /ʊ/ (*hood)* or /ʌ/ (*hud),* /ɔː/ (*hoard)* as /ɑː/ (*hard),* /ɛə/ (*haired)* as /ɜː/ (*heard)* and /əʊ/ (*hoed)* as /ɔɪ/ (*hoyed*), /uː/ (*who'd)* or /ʊ/ (*hood).*

Separate hierarchical cluster analyses for the vowels produced by the LP and HP proficiency groups and identified by the NT listeners were used to analyse the confusion patterns. The resulting analyses are presented in Figures 3.12 & 3.13. Native listeners found some clusters more confusing than others. For the LP group (Fig. 3.12), there were four distinct confusable clusters of vowels: the front vowels, including front closing diphthongs, the high back and low central vowels including high back closing diphthongs, the central vowels, and the back vowels /ɑː/ (*hard)* and /ɔː/ (*hoard).* Within each of these clusters, certain groups of vowels were highly confusable; /ɜː/ (*heard)-/ɛə/ (haired),* /ɒ/ *(hod)*-/ʌ/ (*hud),* /ʊ/ (*hood)*-/əʊ/ (*hoed)*- /uː/ *(who'd),* /eɪ/ *(hayed)*-/aɪ/ *(hide)* and /ɪ/ *(hid)*-/ɛ/ *(head).* The resultant clusters for the HP group in Fig. 3.11, shows some similar patterns. NT listeners frequently confused the high back vowels, /uː/ *(who'd)* and /ʊ/ (*hood),* the central vowels /ɜː/ (*heard)* and /ɛə/ (*haired),* and the high front vowels /ɪ/ (*hid)* and /ɛ/ *(head).* The vowels /ɒ/ (*hod)* and /ʌ/ (*hud)* were somewhat confused, as were /əʊ/ (*hoed)* and /ɔɪ/ (*hoyed).* This latter pair are very different acoustically, and so it is surprising that these were grouped together. One possible explanation is that L2 speakers were not producing these accurately due to orthographic interference.

		Response																	
		iː	ɪ	ɛ	æ	ɑː	ɒ	ɔː	uː	ʊ	ʌ	ɜː	eɪ	aɪ	aʊ	əʊ	ɛə	ɔɪ	Total
	iː	85	0	8	0	0	0	0	0	0	0	5	3	0	0	0	0	0	100
	ɪ	0	46	44	3	0	3	0	0	0	0	0	0	5	0	0	0	0	100
	ɛ	0	10	85	3	0	0	0	0	0	0	0	0	3	0	0	0	0	100
	æ	0	0	0	87	0	5	0	0	0	8	0	0	0	0	0	0	0	100
	ɑː	0	0	0	0	85	3	0	0	3	0	5	0	0	0	0	5	0	100
	ɒ	0	0	0	0	8	33	5	0	15	10	0	0	0	8	21	0	0	100
	ɔː	0	0	0	3	0	0	67	3	3	0	0	0	0	15	5	0	5	100
	uː	0	0	0	0	0	3	3	59	21	0	0	0	0	8	3	0	5	100
	ʊ	0	3	0	3	0	15	3	5	69	0	0	0	0	0	3	0	0	100
	ʌ	0	3	0	8	3	13	0	0	8	49	10	3	0	3	3	0	0	100
	ɜː	0	0	3	0	15	0	3	0	0	0	79	0	0	0	0	0	0	100
	eɪ	0	0	3	0	0	0	0	0	0	0	5	87	3	0	0	3	0	100
	aɪ	0	8	0	0	0	0	0	0	0	0	0	3	90	0	0	0	0	100
	aʊ	0	0	0	0	0	3	3	3	3	0	0	0	0	79	8	0	3	100
	əʊ	0	0	0	0	0	0	0	8	18	0	0	3	0	18	51	0	3	100
	ɛə	0	0	21	8	3	0	3	0	0	0	33	3	3	0	0	28	0	100
Stimulus	ɔɪ	0	0	0	0	0	0	3	0	0	0	0	5	0	10	0	0	82	100

Table 3.5: The confusion matrix showing the percent correct for the vowel intelligibility for vowels produced by the LP group, stimulus in rows, and responses in columns.

		Response																	
		iː	ɪ	ɛ	æ	ɑː	ɒ	ɔː	uː	ʊ	ʌ	ɜː	eɪ	aɪ	aʊ	əʊ	ɛə	ɔɪ	Total
	iː	96	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100
	ɪ	0	41	47	0	0	0	0	0	3	0	0	0	9	0	0	0	0	100
	ɛ	0	21	76	3	0	0	0	0	0	0	0	0	0	0	0	0	0	100
	æ	0	0	6	73	2	0	0	0	0	2	17	0	0	0	0	1	0	100
	ɑː	0	0	1	2	91	3	0	0	0	1	1	0	1	2	0	0	0	100
	ɒ	0	0	0	0	2	28	0	0	47	21	1	0	0	0	0	0	1	100
	ɔː	0	0	0	0	23	8	39	0	1	0	3	0	1	15	2	0	8	100
	uː	0	0	0	0	0	0	0	59	41	0	0	0	0	0	0	0	0	100
	ʊ	0	0	0	0	0	0	0	45	54	1	0	0	0	0	0	0	0	100
	ʌ	0	0	0	12	1	5	0	0	20	62	0	0	0	0	0	0	0	100
	ɜː	2	1	2	0	0	0	1	1	0	0	83	0	0	0	0	10	1	100
	eɪ	0	0	0	0	0	0	0	0	0	1	1	67	31	0	0	1	0	100
	aɪ	0	1	0	0	1	0	0	0	0	0	0	1	97	0	0	0	1	100
	aʊ	0	0	0	0	4	0	3	3	3	4	2	0	0	72	2	0	9	100
	əʊ	1	0	0	0	0	0	0	16	16	1	1	0	1	8	33	0	23	100
	ɛə	0	0	8	2	4	0	2	0	0	0	62	3	10	1	1	8	1	100
Stimulus	ɔɪ	1	0	0	1	1	1	0	0	1	0	1	0	6	1	15	0	73	100

Table 3.6: The confusion matrix showing the percent correct for vowel intelligibility for vowels produced by the HP group, stimulus in rows, and responses in columns.

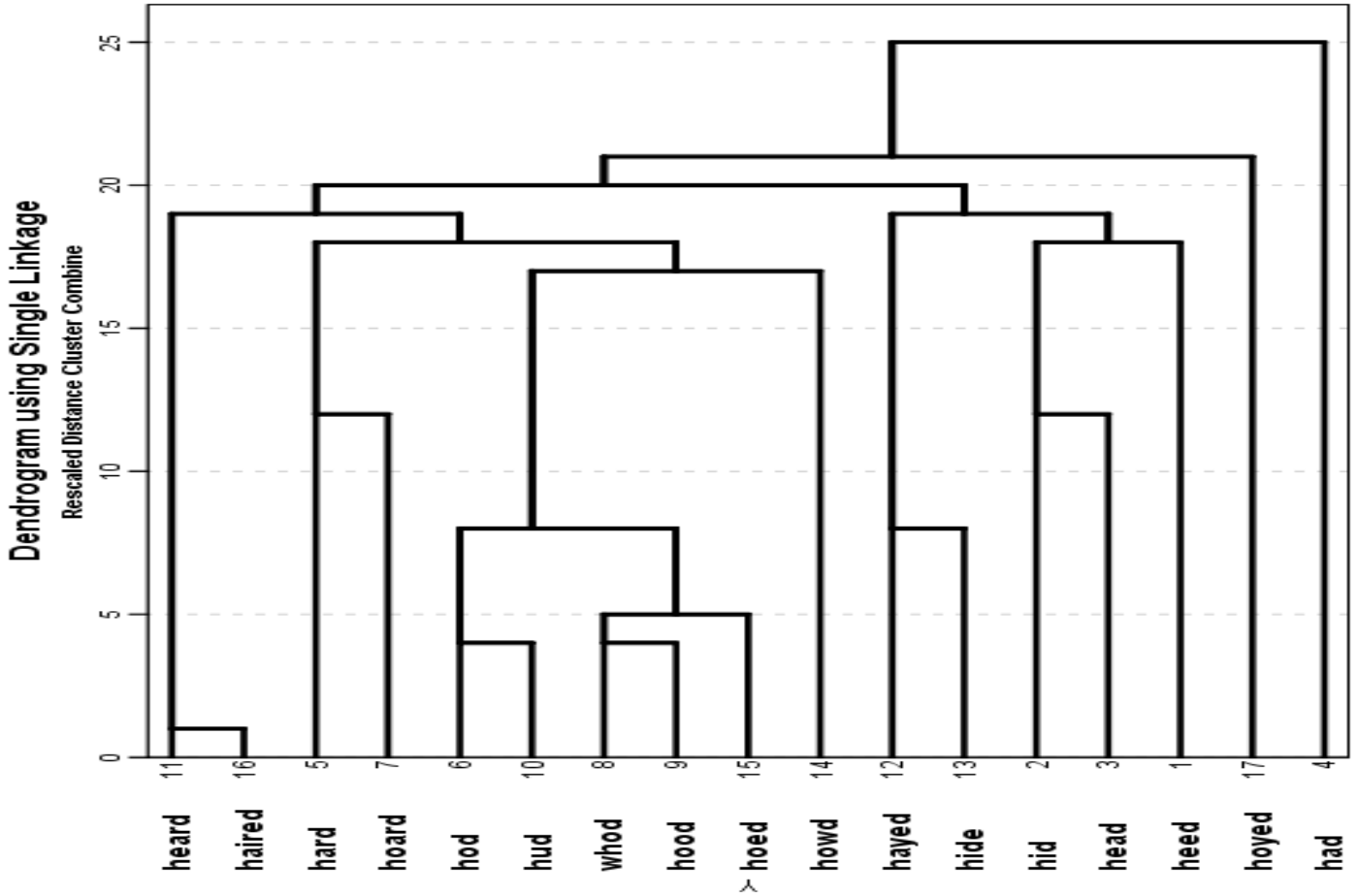


Figure 3.12: Clustering solution showing the distance between the nearest neighbours in the confusion matrix for the LP speakers' vowels as identified by native SSBE listeners.

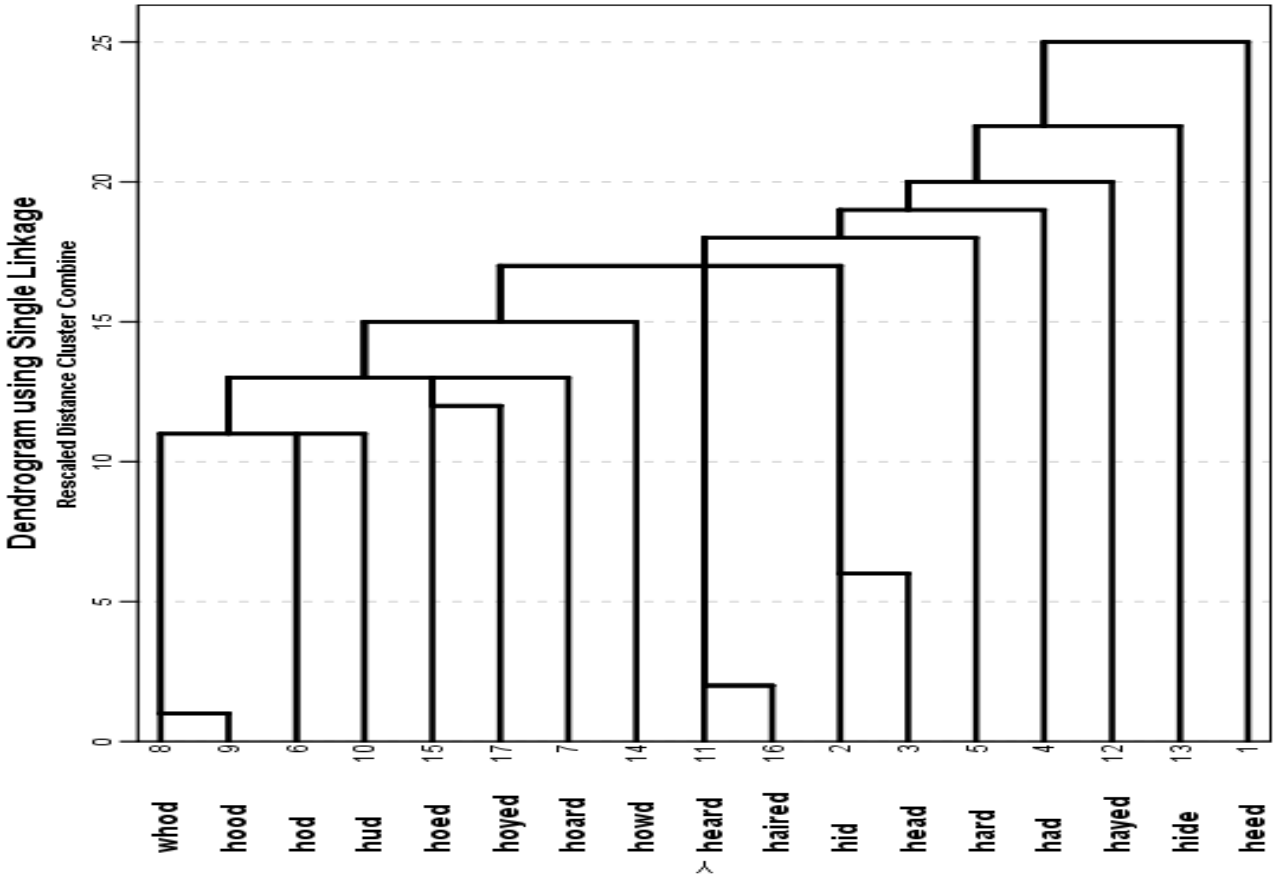


Figure 3.13: Clustering solution showing the distance between the nearest neighbours in the confusion matrix for the HP speakers' vowels as identified by native SSBE listeners.

Accent Ratings

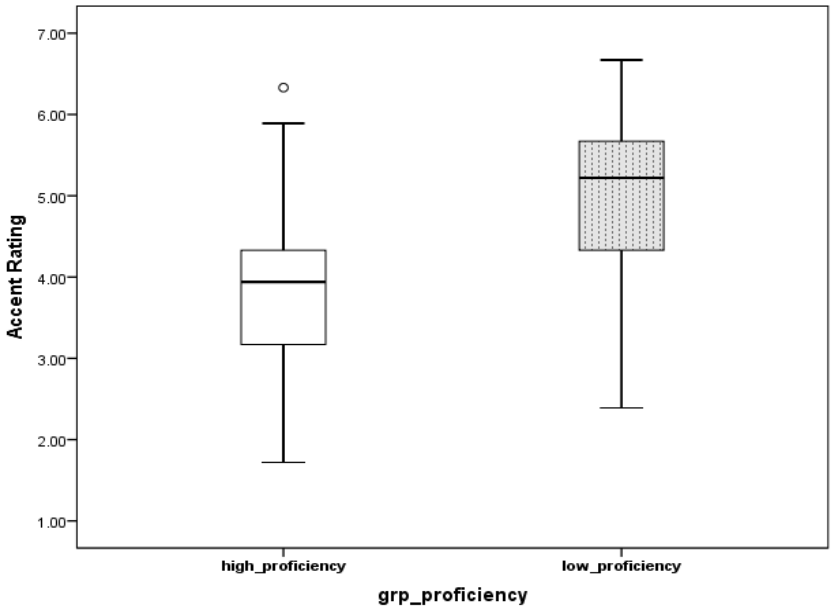


Figure 3.14: Boxplots showing SSBE listeners' accent ratings for L2 Arabic participants' speech. Ratings were made on a scale from 1(native-like) to 7(very non-native).

Before the ratings were examined for differences in performance, it was necessary to establish that the ratings were reliable, i.e., that the raters were using the scale in the same way. A Pearson correlation between all pairs of raters demonstrated that SSBE listeners' accent ratings were in the range of r= .621 to .94, confirming that the ratings had a significant level of agreement. Consequently, an average rating was calculated for each speaker and these values were used in all subsequent analyses.

As displayed in Figure 3.14, there was a large amount of variability in ratings for both HP and LP learners, though HP Arabic learners appeared to be judged to sound more native-like than LP learners. An initial analysis using an independent samples t-test and including all data points indicated that there was no significant difference between groups, p >.05. However, this result appeared to be being driven by the existence of an outlier in the HP group (see Fig 3.14) and an analysis excluding this outlier, demonstrated that there was a significant difference between groups, [t=-2.18, p<.05, df=23].

After examining the difference between the two proficiency groups in the accent ratings, it was of interest to investigate the relationship between vowel intelligibility and accent rating. That is, whether the participants who were rated as more native like tended to be more intelligible than the participants rated as very foreign accented speakers. As displayed in Figure 3.15, there was a significant correlation between ratings and vowel intelligibility, [Pearson correlation, r= -.46, p<.05, R2=.165].

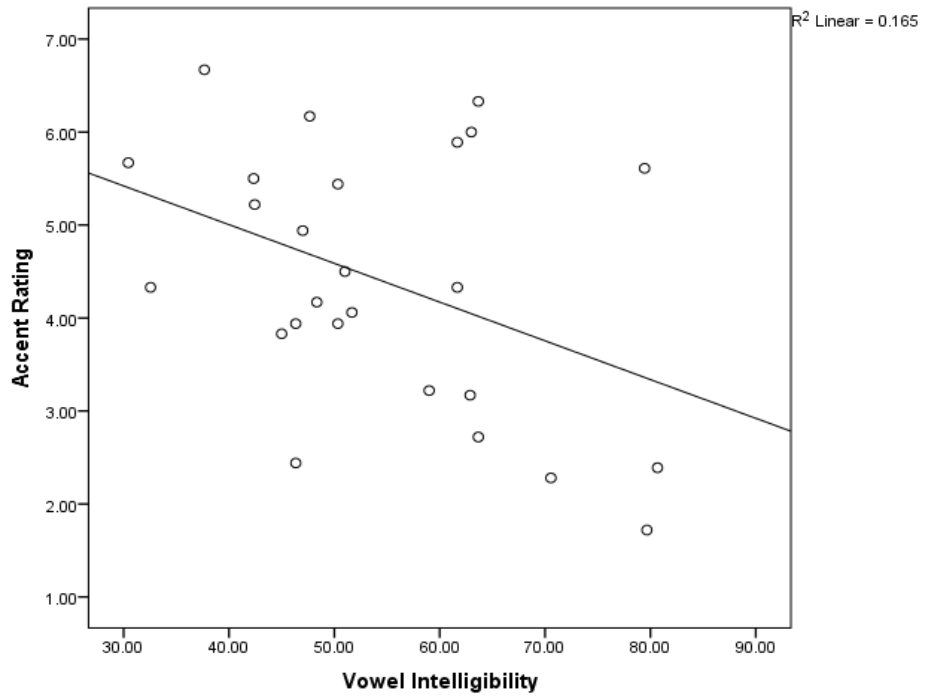


Figure 3.15: Scatterplot showing the correlation between accent ratings and vowel production of Arabic speakers identified by SSBE listeners.

Although listeners may have been basing their ratings on other factors affecting foreign-accentedness, e.g., voice quality and intonation this may suggest that listeners perhaps were paying attention to vocalic features whilst judging foreign accent.

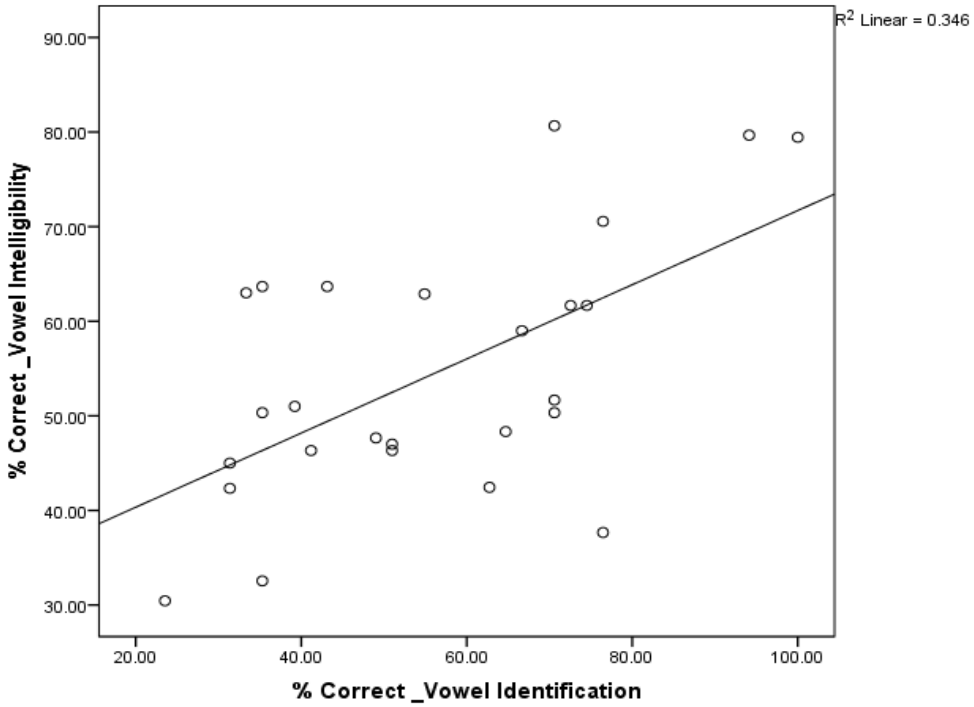


Figure 3.16: Scatterplot showing the correlation between Arabic participants' vowel identification scores and vowel intelligibility.

Figure 3.16 displays the relationship between vowel identification and vowel intelligibility (vowels produced by non-natives and identified by SSBE listeners). There was a significant correlation between vowel identification and vowel intelligibility, [Pearson correlation r=.588, p<.05, R2= .34], indicating performance on a perception task was an indicator of intelligibility.

Informal comparison of the confusion matrices and cluster analyses for vowel identification and vowel intelligibility suggested that groups of vowels that L2 learners found difficult to identify, were also less intelligible. This was particularly noticeable for the LP group. These participants frequently misidentified /ɛə/ (*haired)* as /ɜː/ (*heard)* in the vowel identification task, and their vowel intelligibility of /ɛə/ (*haired)* was also misidentified as /ɜː/ (*heard*). The vowels /ɪ/ (*hid*) /ɛ/ (*head),* and the back vowels /*uː/ (who'd*), /*ʊ/ (hood), /əʊ/ (hoed)* were similarly confused in both perception and production.

Interestingly, perception and production were mismatched for some vowels regarding either the degree of confusion or the change in the confusion pattern. The vowels /eɪ/ (*hayed)* and /aɪ/ (*hide)* were more confusable in the perception than they were in production (closer distance in the perception task than the production). Participants also performed better in production with the vowel contrast /ɔː/ (*hoard)* - /ɑː/ *(hard)* a contrast that they found highly confusable in the perception task. Participants performed badly with /ɒ/ (*hod)* in both production and perception, but in perception they misidentified this vowel as /ɑː/ (*hard),* whilst their productions were misidentified as /ʊ/ (*hood).* Likewise, /ʌ/ (*hud)* was misidentified as /ʊ/ (*hood)* in production but /ɑː/ (*hard)* or /ɒ/ (*hod)* in perception. One possibility is that the orthography may have affected production accuracy for some vowels. For instance, participants might have misread (*hud)* /ʌ/ as *(hood)* /ʊ/, having associated orthographic 'u' with a high back vowel quality, and may not yet have acquired the STRUT vowel /ʌ/.

Similar patterns emerged for the HP group. The vowels /ɪ/ (*hid)* and /ɛ/ (*head*) were problematic in both perception and production tasks, and the vowel /ɛə/ (*haired)* was often identified as /ɜː/ (*heard)* in the perception task, with productions of /ɛə/ (*haired)* were misidentified as /ɜː/ (*heard)*. However, while /uː/ (*who'd)* was confused with /ʊ/ *(hood)* in perception and production, /ʊ/ *(hood)* was confused with /uː/ (*who'd)* in production but not in perception. Similarly, /əʊ/ (*hoed)* was confused with /ɔɪ/ (*hoyed)* in production but not in perception. As for the LP learners, it is possible that even for these more advanced HP learners, orthography may have affected production accuracy.

3.5 Discussion

This study provided initial information about how Saudi Arabic learners of English of varying proficiency levels, perceive and produce the English phoneme inventory. The study used a set of perception and production tasks to investigate the problematic phonemic contrasts for adult Arabic learners of British English. Specifically, the experiments tested whether low and high proficiency groups had difficulty with the perception of the same phoneme contrasts, and how background noise affected the performance accuracy of both proficiency groups compared to

native SSBE listeners. In addition, two production tasks further investigated whether there was a relationship between perception and production accuracy.

3.5.1 Overall performance

The results from the perception and production tasks demonstrated consistent differences between the two proficiency groups in terms of their phoneme identification accuracy in both quiet and noise conditions. For consonants, overall performance for both groups was relatively high (performance on consonant identification in quiet: HP - 82%, LP - 72%), suggesting that as hypothesized, even though both groups experienced some difficulties with consonant identification, learners had for the most part, acquired the English consonant inventory successfully. The HP listeners performed slightly worse than the Arabic-English early bilinguals tested by Shafiro et al., (2012) (82% vs. 95% correct for Shafiro et al.'s study). It is also important to note that these results demonstrated that HP learners confused /dʒ/ and /ʒ/, a contrast omitted in the Shafiro et al (2012) study.

As expected, overall performance in the vowel identification task was harder than that of the consonants for both groups; HP listeners scored 68% whilst LP listeners scored 46%. The reduced vowel accuracy compared to consonant accuracy correct may suggest possible effects of the difference in phonemic inventory between Arabic and English. In vowel identification Arabic listeners were presumably mapping the larger English vowel inventory to a small Arabic vowel system, as has been shown in previous studies (Iverson and Evans, 2007; Escudero and Boersma, 2002; Shafiro et al., 2012), whereas there were more possibilities for direct one-to-one mapping across the two consonant inventories. Iverson and Evans (2007) trained German and Spanish learners of English on the perception of English vowels, and found that German speakers who have 15 monophthongs, and three diphthongs in their L1 vowel inventory benefited more from the vowel training than the Spanish speakers who have 5 vowels in their L1 vowel inventory. This suggests that a large L1 phonemic inventory may facilitate L2 phoneme perception/learning whilst a small inventory, like that of Arabic, may make learning more difficult, at least initially.

3.5.2 Error patterns

Both HP and LP groups had some confusions in perception task, and the same confusions occurred in the vowel intelligibility task. Despite differences in overall proficiency, the error patterns for both groups were remarkably similar. Thus, although HP learners had become better at identifying English consonants and vowels, (perhaps through greater experience with English they had developed more detailed category representations), their processing was still affected by their native language.

For the consonants, both groups experienced difficulties with the English voiced affricate /dʒ/ which was identified either as the postalveolar fricative /ʒ/ or the plosive velar /g/, and the postalveolar fricative /ʒ/ which was either identified as the voiced affricate /dʒ/ or the plosive velar /g/. Interestingly, early English-Arabic bilinguals in Shafiro et al.'s study (2012) experienced fewer difficulties with /dʒ/, though they did also confuse this with /g/. However, listeners in the Shafiro et al. study were not tested in their perception of /ʒ/. One possible explanation for our results is the effect of dialect background. Although the phoneme /dʒ/ exists in the MSA, Saudi speakers use the variant /ʒ/ in their low variety in place of /dʒ/. Our pilot study (see p. 40 for summary) showed that Saudi speakers use the voiced affricate /dʒ/ in formal settings like reciting the Qur'an, but in informal settings they use the variant /ʒ/ instead. This indicates that /ʒ/ and /dʒ/ may be allophonic variants and may thus both be assimilated into the same underlying native category (e.g., a single category assimilation) according to the PAM model. That is, when Saudi speakers hear the English phonemes /ʒ/ and /dʒ/, because they are close to one L1 phoneme /ʒ/, they assimilate the two phonemes to the one L1 category that they use mostly in their L1 /ʒ/. This also suggests that phoneme categorization may be highly specific, and that L1 dialect may play a significant role in L2 perception.

This in line with previous work showing that Czech listeners from different dialect backgrounds show different patterns in their acquisition of Dutch monophthongs (Chla´dkova´ and Podlipsky, 2011). Czech listeners from Bohemia and Moravia were tested in their perception of 12 Dutch (western part of Netherland) monophthongs presented in /h/-V-/b/ nonsense words that were displayed using Czech orthography. The vowels are differentiated by their spectral properties; eight of these

vowels are phonetically short, and four are long. However, the Czech spoken in Bohemia and Moravia has ten monophthongs; five short and five long. Though many of the vowels in both dialects share similar vowel spectral properties and duration, the high front vowel contrast /iː-ɪ/ differs between the dialects such that in Bohemia, these vowels are distinguished perceptually more by spectral differences and by smaller duration difference than in Moravia (Podlipsky et al., 2009). Moravian but not Bohemians favour the durational differences over the spectral when perceiving the Czech /iː-ɪ/, and the Dutch /i- ɪ/ contrast is realised by spectral properties. Chla´dkova´ and Podlipsky (2011) found that the two groups had different assimilation patterns for Dutch high front vowels; since the Dutch vowel /i/ is short, the Moravian Czech listeners assimilated both Dutch vowels /i-ɪ/ to a single native category /ɪ/. However, Bohemian Czech listeners perceived the Dutch /i-ɪ/ vowel contrast in terms of their two native categories /iː-ɪ/. Chla´dkova´ and Podlipsky suggested that even slight acoustic and therefore perceptual differences between individuals' native dialect can affect L2 speech perception.

In the current study we tested Saudi Arabic participants from two different areas, Riyadh and Jeddah so the participants may have had different dialect backgrounds. However, informal examination of our results suggested that dialect background did not affect perception of L2 phonemes in our listeners. One reason for this is that, although both dialects use the /dʒ/ and /ʒ/ differently (Riyadh speakers use /dʒ/ variant in spontaneous speech and in formal settings, while Jeddah speakers use it only in the formal settings), they both contain these variants. Another reason could be that a large number of the Riyadh participants had close contact with Hijazi speakers, and so may have been highly familiar with the dialectal variants in both varieties, and use both in daily conversation. One possibility then is that participants in the current study confused this contrast (i.e., /dʒ/ and /ʒ/), perhaps because they use it in their L1 interchangeably, and this interchangeable use does not affect their intelligibility in their native language.

Listeners also had difficulties with the voiceless affricate /tʃ/. Unlike /dʒ/ this does not exist in the Arabic consonant inventory (Appendix 1), and so one might expect listeners to assimilate this phoneme to the nearest native category, /ʃ/. Both HP

and LP listeners displayed this pattern, (i.e., assimilate /tʃ/ to L1 /ʃ/). However, it was notable that HP listeners performed much better with the voiceless affricate (/tʃ/ - 76%, /dʒ/ - 59%) than LP listeners who performed similarly for /dʒ/ and /tʃ/ (31% and 33% respectively). This pattern of results suggests that Arabic learners found it easier to acquire the voiceless rather than the voiced affricate. One could imagine that this is because learners found it easier to acquire a sound outside their native consonant inventory, rather than adjusting their underlying phonological representations (i.e., learning that /dʒ/ and /ʒ/ are separate phonemes rather than allophones of /dʒ/).

Both HP and LP listeners had difficulties in identifying the velar nasal /ŋ/ which was most frequently misidentified as the alveolar nasal /n/. This is probably because the phoneme /ŋ/ does not have a counterpart in Arabic and therefore Arabic listeners assimilated it into the phoneme /n/ which is the closest Arabic consonant to the English velar nasal. It should be noted though that both native and non-native listeners had difficulty identifying /ŋ/ in noise. Whilst we did not test native speakers in quiet, it is possible that this phoneme may be difficult to identify in these stimuli (VCV) rather than being a result of L2 category assimilation.

Remarkably, listeners had few difficulties with /p/ despite the fact that in Arabic, there is no equivalent to the English phoneme /p/. Previous research has shown that Saudi Arabic speakers confuse /p-b/ in production, producing /p/ with a VOT similar to that of native /b/ (Flege and Port 1981). Surprisingly, Saudi subjects in the current study identified /p/ fairly accurately; 86% for the HP and 74% correct for the LP group. This finding mirrors that of Shafiro et al. (2012) who reported that early Arabic bilinguals also did not experience difficulties with the English phoneme /p/. One possible explanation is that perception and production operate differently (e.g., Evans & Iverson, 2007; Hattori & Iverson, 2010). Thus, although learners may have had a non-native like production, this phoneme may have been uncategorizable and therefore, easier to perceive (cf. Best et al., 2001).

For vowels, those that do not have counterparts in Arabic, (/ɛ/ (*head),* /ɜː/ *(heard),* /ɛə/ *(haired),* /ɒ/ (*hod),* /ʌ/ (*hud),* /əʊ/ (*hoed),* /uː/ (*who'd),* /ʊ/ (*hood)*) were found to be more difficult for Arabic listeners to identify. The reason for confusing

these vowels could be that Arabic learners of English were mapping the more complex English vowel inventory onto their smaller Arabic vowel system. For instance, both HP and LP groups had difficulties with the monophthongs /ɒ/ (*hod)* and /ʌ/ (*hud),* and the diphthongs /ɛə/ *(haired)* and /əʊ/ (*hoed)*. However, at least for some of these vowels, it is possible that orthography may have affected identification performance. For instance, /ɒ/ (*hod)* was misidentified as /əʊ/ (*hoed)* and /ʊ/ (*hood)* which are very different acoustically, but have similar orthography and which L2 learners may have associated with same pronunciation. It is possible that even though the responses included familiar rhyme words, participants were not familiar enough with the stimuli to be able to use this information effectively, particularly during this task.

The results also provided evidence that learners with high proficiency in English had acquired new vowel categories. Arabic does not have the high front vowels, /ɪ/ or /e/, though it does have /i/ and uses duration contrastively. Consequently, it was expected that the duration cue would help them distinguish between /i/ and /ɪ/, but since the duration would not help in distinguishing between /ɪ/ -/e/, that Arabic learners would have difficulty with English, /ɪ/ and /e/. LP participants consistently misidentified /ɪ/ (*hid)*, as /e/ (*head)* rather than /i/ *(heed).* This indicates that they were able to transfer their use of duration as a cue and that they had started to establish a new category midway between their native /i/ and /a/ which they used for English /ɪ/ (*hid)* and /e/ (*head)*. However, HP learners had started to further split the acoustic space; they did misidentify /ɪ/ (*hid*) as /e/ (*head)* but not to the same extent as the LP learners. This contradicts the findings of Shafiro et al. (2012) who found that these high front vowels were the least confusable. This is probably because their Arabic participants were highly proficient (early bilinguals).

All listeners had difficulties with central (/ɜː/ (*heard), /eə/ (haired)*) and high back vowels. This could be explained by the fact that Arabic only has a single high back vowel /u/ and no central vowels. It is likely then that listeners assimilated all English back vowels into their single back vowel /u/. This mirrors similar patterns of assimilation with other L2 groups with a similar L1 space, for example, Spanish learners who have small vowel system (only 5 vowels; /i/, /e/, /a/, /o/, and /u/) found English vowels /ɒ/-/ɔ/ both sounded like /o/ in Spanish and hence misidentify both

English phonemes as their L1 phoneme. (Iverson and Evans, 2009). It is less clear why central vowels were confused. Despite the fact that these two vowels do not exist in Arabic, one might have expected that Arabic listeners would be able to identify these vowels as distinct categories that are not a good fit to either of their nearby Arabic categories (i.e., /iː/, /i/, /aː/, /a/). However, given the fact that they confused these two central vowels with each other (/ɜː/ (*heard)*, /*eə*/ *(haired)*), but not with either of their L1 vowels, this may indicate that they could recognise them as different from their L1 vowel inventories, but still their recognition was not robust enough to distinguish these central vowels as two separate vowel categories, perceiving them as a single vowel category. This may also be due to acoustic factors; /*eə*/ *(haired)* has very little formant movement and its onset is similar to that of the central vowel /ɜː/ (*heard)*.

3.5.3 Effect of noise on vowel and consonant identification

As expected, accuracy of both vowel and consonant identification decreased as the noise level increased, for all participants. In both vowel and consonant identification in noise, native SSBE listeners performed better than the Arabic listeners which mirrors findings from other studies (e.g., Cooke et al, 2008). As predicted, there was a difference between the two proficiency groups' performance in different noise levels; the HP group tended to perform better than the LP group, confirming that less experience with an L2 leads to more difficulties in comprehension in noise for L2 phonemes.

Noise affected the identification of vowels and consonants differently. Previous work by Cutler et al. (2004) showed that Dutch listeners' identification of English vowels was not greatly affected by noise, but that identification performance for consonants was poorer in their lowest noise condition (0dB SNR). This could be because in Cutler et al.'s study the SNRs were higher (i.e., less noise and easier to understand; 0, 8, 16 dB) while in this study the SNRs used were much lower (i.e., more noise and harder to understand 0,-5,-10 dB). Additionally, Cutler et al used babble noise rather than speech-shaped noise, as was used here. Another possibility is that as Cutler et al.'s participants were Dutch and Dutch has a more complex vowel space, their participants were able to rely on direct mapping between Dutch and English vowels. In contrast, our Arabic listeners found vowel identification in noise harder

than consonant identification, which was expected given that Arabic listeners performed worse in vowel identification in quiet than the consonant identification in the quiet condition. Furthermore, whilst noise affected consonant identification for all subjects equally, non-native listeners' vowel identification performance was more affected at lower noise levels. Arabic listeners, even those who perform well in quiet, may not be able to rely on such strategies (being unable to map to native categories) which may mean that they are reliant on less well defined categories which break down more easily in noise.

However, in comparison to the native SSBE listeners, non-native speakers' performance did not drop dramatically as the noise level increased; the difference in performance at each noise level was bigger in SSBE listeners than in that of the non-natives. This contradicts what Cutler et al (2004) found; that the performance asymmetry between native and non-native listeners was not different across different SNRs. This again is perhaps because Cutler et al.'s study used easier levels of noise, and different noise masker (babble noise). Another possible reason is that the Arabic participants in our study performed poorly at the easiest noise level (0dB), so increasing the noise dropped their performance, but not in as dramatic a way as for natives' performance (see Fig 3.10).

3.5.4 Production-perception link

This study only investigated the relationship between the perception and production of vowels. There was some evidence for a link between production and perception (cf. Bradlow and Pisoni, 1996). Accent ratings and vowel intelligibility (i.e., SSBE listeners' identifications of Arabic participants' vowels) were significantly correlated; Arabic participants who were given more native-like ratings were also more intelligible. Vowel identification and vowel intelligibility were significantly correlated and there were also similarities in the error patterns in production and perception. That is, the same problematic vowel categories in perception were found to be problematic in production (e.g., /uː/ (*who'd)*, /ʊ/ (*hood)*, and /ɪ/ (*hid)*).

However, there were some differences and vowel categories which were not confusable in perception, were found to be confusable in production, (e.g., /ɔː/ (*hoard)*-

/*a:*/ *(hard)*). Furthermore, there was no correlation between vowel identification performance and accent ratings. One possible reason for this is that factors such as voice quality and prosody might have affected ratings but did not affect intelligibility. That is, SSBE listeners might have found a speaker highly intelligible, but based their accent ratings on factors other than intelligibility. Equally, it is possible that if recordings that included more examples of the problematic vowel categories had been used, there may have been a positive correlation between identification and accent ratings (cf. Hattori and Iverson, 2009). Hattori and Iverson (2009) included an accent revealing sentence that included the notoriously problematic consonant contrast for Japanese learners of English /r/-/l/ "*The red robin looked across the lovely lake*". In this study, it was not clear at the outset which vowels would be problematic for Arabic learners of English and so a sentence that included both vowels and consonants that were expected to be difficult, including a consonant cluster which is not permitted (LP Arabic learners of English epenthesize a vowel between a consonant cluster to break it) in Arabic was chosen; "*Then the North Wind blew as hard as he could, but the more he blew, the more closely did the traveller fold his cloak around him; and at last the North Wind gave up the attempt*".

3.5.5 Summary

The current study explored problematic vowel and consonant contrasts for Saudi Arabic learners of English. As expected, the contrasts that do not occur in Arabic presented the most difficulty for the learners. In particular, Arabic learners had difficulties with English affricates, high front, and high back and central vowels. In contrast to previous work (e.g., Cutler et al., 2004), all Arabic listeners, regardless of proficiency, found vowel identification harder than consonant identification in both quiet and noise. Additionally, the study provides some evidence for a link between perception and production; perception of English vowels was better in Saudi learners who also had more accurate production of these vowels. The next chapter uses these results as the basis for a training study that investigates the relationship between production and perception in more detail.

Chapter 4 Introduction to Chapters 5-7: Investigating the relationship between speech perception and production

4.1 Overview

In the previous chapter it was shown that Arabic learners of English find certain English vowel contrasts more challenging in both perception and production than English consonants. Additionally, measures of vowel perception and production were correlated, suggesting that accurate perception and production of vowels is linked in some way. This chapter explores the links between speech production and perception in light of previous studies of L1 and L2 acquisition (see chapter 2 for review). The aim of exploring the link between perception and production in this chapter is to present the rationale behind the training study presented in Chapter 5 in which Arabic learners of English were trained in their production and perception of English vowels in 3 different training conditions; production-based training, perception-based training, and a hybrid training condition that gave training in both perception and production.

Of interest, was whether training in one domain would generalise to the other untrained domain. That is, if participants completed perception-based training, for example, would this lead to improvements in production as well as perception?

4.2 Introduction

4.2.1 The relationship between perception and production: do changes in perception lead to changes in production?

The relationship between speech perception and production has been a long-standing focus in experimental phonetics and speech science. Several theories of speech perception have claimed a strong relationship between perception and production (e.g. Liberman et.al, 1989), suggesting that perception and production share common underlying representations. Probably the most well-known of these approaches is Motor Theory (Liberman et al., 1967, Liberman and Mattingly, 1985; Galantucci et al., 2006) which postulates that listeners perceive speech though articulatory gestures. When perceiving speech, listeners are thought to access their

own knowledge of the way that phonemes are articulated. Specifically, according to the Motor Theory, listeners perceive speech as the speaker's intended articulatory gestures (e.g., the intended movement of tongue, lips and jaw raising), but not the actual acoustic patterns generated by the articulatory gestures. Motor Theory thus hypothesizes that individuals perceive speech with a speech-specific system or module (i.e., phonetic module), but not with a general perception mechanism. For speech perception, the phonetic module detects the intended articulatory gesture (i.e., the neuromotor commands that call for movements of the articulators through certain linguistic configurations) from the acoustic signal, and then relates such information to abstract phonological knowledge. For speech production, the module translates the abstract knowledge to the neuromotor commands to produce the intended realization of phonemes.

Similarly, Direct Realism (e.g., Fowler, 1981, 1986; Best, 1995) claims that the objects of speech perception are articulatory rather than acoustic events. Direct Realism argues that the information in the acoustic pattern (i.e., the waveform) is sufficient for the individuals to specify the actual articulatory gestures (e.g., lip, tongue and jaw movements), and that a listener reconstructs a speaker's actual articulatory movements via the acoustic wave form that is shaped by the speaker's articulators. Unlike Motor Theory, Direct Realism does not presuppose any special mechanisms corresponding to the phonetic module, rather, it hypothesizes that individuals use general perceptual systems, which have a universal function and include the same means by which animals can perceive or know the environmental conditions in which they live. Even though Direct Realism suggests that the perceptual systems use acoustic structure (the waveform) that is caused by the articulatory movements (e.g., lip movements) as information for the movements, it is not the waveform that individuals perceive, but the actual articulatory gestures.

The General Auditory Approach (GAA; Diehl et al., 2004, p.167) also suggests a very close relationship between speech perception and production. GAA argues that perception follows production and production follows perception by offering two general accounts of speech processing. The first assumes that the auditory distinctiveness of phonemes shapes production. That is, if a speaker speaks clearly in

a situation that demands clear auditory characteristics, the speaker may maximise interphoneme distance in the phonetic space to promote intelligibility (e.g., increasing the F1 value for English /e/ to distinguish it from English /ɪ/). According to the GAA the perceptual demand in the clear speech "sharpens up" the speaker's production, and thus speech production follows speech perception.

The second account proposes that perception follows production. According to the GAA, listeners perceive the acoustic consequences of gestures. That is, any regularities of speech production will be reflected in the acoustic signal and it is these regularities which listeners access when comprehending speech. Listeners are thus, thought to make use of the acoustic correlates of production regularities in judging the phonemic content of speech signals (see Diehl et al., 2004 for review).

Other supporting evidence for the link between speech perception and production comes from brain imaging studies (e.g., Rizzolatti and Arbib, 1998; Fadiga et al, 2002; Wilson et al., 2004). For example, Wilson et al. (2004) used fMRI to examine whether the motor areas that are involved in producing speech are activated when listening passively to meaningless monosyllables. Wilson et al tested 10 participants who listened to 16-blocks of stimuli whilst being scanned, each containing 23 repetitions of meaningless monosyllables. During these scanning sessions, participants were asked to produce the same syllables. Wilson et al found that for all participants, there was substantial overlap when comparing the regions activated by listening to and producing the syllables. These findings are consistent with the view that speech perception involves the motor system in a process of auditory-to-articulatory mapping to access a phonetic code with motor properties.

However, the link between speech production and perception from behavioural studies is not as clear-cut. Although some studies of adult second language learning have shown that perceptual training leads to improvements in both speech perception and production (e.g., Bradlow et al., 1997; Yamada et al, 1995; Wang et al, 2003; Hazan et.al, 2005), other studies have found little or no relationship between perception and production (e.g., Bailey and Haggard, 1973, 1980; Ainsworth and Paliwal, 1984; Hattori, 2009; Hattori and Iverson, 2009). For example, Wang et al

(2003) trained American English speakers on Mandarin tone perception and after eight training sessions of 40 minutes completed over a two week period, learners had improved not only in their tone perception, but also in their tone production.

In contrast, Bailey and Haggard (1980) found a weak relationship between young children's perceptual category boundaries for a /k/-/g/ continuum and average VOTs produced in voiced and voiceless consonants in their L1. They tested 34 children (average age 3 years-old) on the perception and production of five initial voiced-voiceless contrasts (*bin-pin*, *deer-tear*, *goat-coat*, *girl-curl*, and *bear-pear*). The results demonstrated that there was no correlation between average VOTs produced for voiceless and voiced consonants and listeners' perceptual category boundaries for a /k/-/g/ continuum. Similar patterns of results have been found for children acquiring an L2. For example, Tsukada et al. (2005) tested Korean children in their discrimination and production of English vowels, and found that Korean children were better at producing English vowels than they were at discriminating them. That is, they produced vowels that were as intelligible as those of native age-matched English speakers, but did not perform as well on a vowel discrimination task as these native speakers. Similarly, for adult learners, Sheldon and Strange (1982) demonstrated that some Japanese speakers were more accurate at producing the English /r/-/l/ contrast than they were at identifying it (see also Goto, 1971).

More recently, Hattori (2009) investigated whether training in production rather than perception would lead to improvements in production in L2 learners, and whether or not this learning would generalize to perception. Twenty-eight native Japanese speakers with varying levels of experience with English were trained in the production of English /r/-/l/ in ten 30-40 minute sessions completed over a 2-3 week period. The training combined three methods. First, participants were given explicit feedback and instructions through one-on-one interactions with a phonetics teacher. For example, listeners were taught where to position their tongue to produce /r/ and /l/, used a mirror to monitor their own tongue positions, and watched video recordings of model /r/ and /l/ productions to observe tongue movements. Secondly, participants were shown real-time spectrographic displays of their speech so that they could visually monitor their formant frequencies as they spoke. Thirdly, participants' own

productions were signal-processed to make them closer to native-like pronunciations of /r/ and /l/ (e.g., Iverson et al., 2005), so that individuals could compare their own speech to an idealized target spoken in their own voice. Hattori (2009) found that after ten sessions of the pronunciation training Japanese speakers were able to produce native-like /r/-/l/, but that their perception of this consonant contrast was not improved. These findings suggest that speech perception and production might not share similar underlying representations or at least, that the learning mechanisms for L2 speech perception and production operate somewhat independently.

The current study further investigates the relationship between perception and production by comparing the results of three training approaches for the acquisition of British English vowels by native Arabic speakers. The first approach used a one-to-one production training paradigm based on articulatory phonetics. Recently, pronunciation has been taught using different means. For example, as described in the previous paragraph, Hattori (2009) used real-time spectrograms to display Japanese speakers' production of English /r/ and /l/, so that they could monitor their speech visually. Participants were given training on how to interpret variation in F3 so that they could pay attention to the acoustic consequences of the articulatory movements crucial in distinguishing /r/ and /l/ and compare their spontaneous speech with signal-processed versions of /r/-/l/ based on their own voice but changed to sound like that of native speakers. This kind of spectrographic feedback has also been successfully used in clinical studies (e.g., Chaney, 1988; Hagiwara et al., 2002; Huer, 1989).

Despite the success of using spectrogram feedback in adjusting the pronunciation of a single consonant or consonant contrast, it would arguably be challenging to use for vowel production training. In cases such as those described above for the training of /r/ and /l/, participants could be directed to a single feature (i.e., the third formant value). Crucially, participants did not need to understand a great deal about spectrograms or be taught in detail how to read them; they just needed to know how to look for and recognise a particular feature which occurred in a given position (e.g., word initial position). Vowel training has been shown to be less successful when sub-sets of vowels contrasts are trained (e.g., Nishi and Kewley-Port, 2007; see p. 33) and so in this study, participants were trained on ten English

monophthongs and four diphthongs, covering the majority of the vowel space. Learners would thus have needed to learn a far larger number of spectrographic patterns in order to be able to identify each vowel, and additionally, would have needed to learn how to compare the formants of each vowel with that of a native speaker's or their own signal-processed recordings.

Instead of using spectrograms for teaching pronunciation, several studies have trained L2 learners on pronunciation using automatic speech recognition (ASR) based computer-assisted language learning (CALL) systems (e.g., Dalby and Kewley-Port, 1999; Yamada et al., 1998; Chou, 2005; Neri et al., 2008). For example, Dalby and Kewley-Port (1999) developed a CALL system, PRONTO, for native speakers of American English learning Spanish, and for Mandarin Chinese speakers learning English. In PRONTO, participants identify minimal pairs spoken by different talkers (word identification), and repeat words presented aurally (word imitation). Their response is then evaluated by the recognizer. They then respond to visually presented prompts by speaking the word with no immediate auditory model (word production). The PRONTO system records the performance continuously on each of these tasks, evaluates it, and then gives feedback displayed to the instructor and the student in a bar chart, that shows the performance level for the perception and production for the minimal pairs. In addition to keeping score for each task for each minimal pair, the system keeps a global score that sums the scores by task. Participants may choose which minimal pairs and which task they wish to practice, but their overall intelligibility profile will improve more if they show improvement on the phonetic contrasts that are more highly valued by the global training score (minimal pairs on the bar chart are listed from top to bottom according to their importance). It is however, unclear how useful such a system would be for L2 learners. Though PRONTO compared to other systems at the time, no formal testing was carried out with L2 learners.

Neri et al (2008) developed an ASR-based CALL system called Dutch CAPT (Computer Assisted Pronunciation Training) that assesses pronunciation and gives automatic feedback either in Dutch or in English, on Dutch pronunciation in various speech styles. The programme gives feedback on the pronunciation of eleven Dutch

phonemes that have been found to be problematic for Dutch L2 speakers from different L1 backgrounds (see Neri et al., 2004). The ASR module analyses the spoken word by looking for the problematic phonemes. The L2 learners took part in role-plays, and answered questions by pronouncing one of the possible answers. They also produce a set of minimal pairs for each contrast. If the participant pronounced the word correctly, an orthographic representation of the utterance they had produced along with a smiley face and short comment was displayed on the screen. However, if the ASR algorithm detected a phoneme that had been mispronounced, an orthographic representation of the utterance was displayed on the screen with the corresponding letter(s) coloured red and a red disappointed face with a comment/ message informing the participant that the red sound(s) had been mispronounced. The participant was then prompted to repeat the utterance. The results indicated improvement in pronunciation, but this was not significantly different from a control group who had no training.

CAPT systems have also been combined with phonetically-based approaches to pronunciation training. For instance, Wik (2011) developed a virtual language teacher (VLT) as a vowel-learning tool for L2 learners of Swedish from different L1 backgrounds. The main focus of the VLT software is a 3D canvas with a ball, and a vowel chart that corresponds to the vowel uttered by a speaker. This gives immediate feedback on the consequences of the speaker's articulatory movements. The learner is prompted to produce a given phoneme in isolation and, may choose from two modes, a practice mode and game mode. In practice mode, the learner is free to choose a vowel to practice on, with no time restrictions; the learner can click on the chosen vowel by clicking on a button and, the corresponding target sphere will appear on the canvas. When there is no sound input, the moving ball will return to the neutral position in the centre of the canvas. In game mode, which is a 'catch-the-target-sphere' race against time, the target spheres are placed on the vowel chart, one at a time, and remain there until the learner manages to keep the moving ball steadily inside the target sphere for 500ms. The target sphere then turns to green, and is replaced by a new one at another position, corresponding to another vowel. Given that the task was to keep the moving ball in the target sphere for at least 500ms, Wik used only long vowels.

Wik (2011) suggested that the immediate visual feedback given by the moving ball helps an L2 learner to discover the relationship between configuration of tongue, mouth and positions on the vowel chart. That is, by moving the tongue backwards or forwards the ball will move from right to left or left to right respectively on the canvas. Similarly, by opening and closing the mouth the ball moves up and down respectively on the canvas. Wik found after training L2 learners over two sessions, that there was some evidence of a learning effect when comparing the performance in the first and the second sessions, however, this was a small effect.

In spite of the apparent advantages of the CALL system in helping L2 learners to improve their pronunciation (e.g., immediate feedback, no need for a teacher), there appears to be some drawbacks, and some learners do not appear to benefit greatly from this kind of training. One possible reason why Neri et al (2008) and Wik (2011) failed to find convincing improvements as a result of production training is perhaps because they trained learners from a range of L1 backgrounds. Participants may therefore have had different degrees of difficulty with acquiring English consonants and vowels. Additionally, there may have been some drawbacks with this approach in general. One drawback is the form of the feedback which might not be helpful for some learners, in that it does not tell the L2 learner why their pronunciation of certain phoneme is close or far from the native speaker's (good or bad). Another drawback is that even if the L2 learners know that their pronunciation is incorrect, in order for the learner to learn the native pronunciation, they need explicit feedback on their mispronunciation (e.g., articulatory feedback or clear visual feedback).

One way in which a number of studies have tried to overcome these drawbacks is by using Virtual Talking Heads (VTH) such as Baldi (Cohen and Massaro, 1995; Massaro and Light, 2003, 2004; Massaro et al., 2011), MASSY (Fagel and Madany, 2008) and ARTUR (Engwall and Bälter, 2007; Engwall et al., 2006). For example, Massaro and Light (2004) found that children with hearing loss improved in their performance on various speech perception and production tasks after completing number of training sessions with Baldi. Children with hearing loss aged 8-13 years-old completed two training sessions per week over a course of 21 weeks, including 2 weeks break on 8 problematic categories including the distinction between voiced vs.

voiceless contrasts /f-v/, /θ-ð/, /s-z/, /t-d-b/ (/p-b/ was not necessary as instructors indicated), consonant clusters [two consonant word initial clusters including /r/ (e.g., *cry, grow ,free*) and /s/ (e.g., *smile, slit, stare*), and two-consonant word final clusters involving /l/ (e.g., *belch, milk, field*)], and the fricative versus affricate distinction /ʃ/-/tʃ/. Baldi speaks slowly, and has a transparent skin that reveals the articulators, including the tongue, teeth and palate as well as fold. During training each participant completed two sessions a week where they were trained on both perception (identification task) and production (listen and repeat isolated words). Feedback was given after each trial (a happy or sad face representing a correct or incorrect responses respectively appeared on the monitor). For the production task, feedback was given as judged by the experimenter (the experimenter input to the computer after each response determines the feedback). Children also completed pre- and post-tests in which they listened to and repeated isolated words produced by Baldi that included all the training segments in all contexts. The results showed that after training, the children improved in speech perception and production, and that improvement in production generalized to new words. However, improvements as a result of training did not appear to be retained; in a follow-up test completed 6 weeks after training where production had deteriorated.

Such talking heads have also been used to train L2 learners. Massaro and Light (2003) used Baldi to teach Japanese learners of English the /r/-and /l/ contrast. Learners were trained using one of two models; front-facing 'normal' view of Baldi (i.e., no internal articulators showing), and a view of Baldi that also showed internal articulatory processes in the oral cavity. The results showed that both types of training were effective, but that interestingly, those who were trained with the view of Baldi showing the articulators, did not improve significantly more than those shown the 'normal' view. Massaro and Light suggested that this was because there were only 11 participants, for two of the three training stimuli there were ceiling effects, and participants had only 3 training sessions, which might not have been sufficient for the learners to master the remaining contrast.

Similarly, Engwall (2008) used a computer-animated virtual teacher (ARTUR) to teach seven French participants the pronunciation and articulation of Swedish

words. The participants were trained on the Swedish words; (*rik*, *rak*, *kora*, *kir*, *karikerar*, *schack*, *sjuk*, *skick*, and *chock*). These words were chosen because they contain the phonemes [r] and [ɧ] in different vowel contexts, but with [k] word-finally. During training, acoustic and articulatory data were collected using a set-up including audio and video recordings, an ultrasound scanner and electromagnetic tracking system. Each participant's attempt for each word was recorded separately. Articulatory data were collected using a Logiq5 ultrasound. A hand-held transducer (ultrasound probe) was used to give the participants the opportunity to identify any changes in the articulation between different attempts, and a tracking system was used to monitor whether the participant was holding the transducer correctly in the midsagittal plane.

All instructions were given by ARTUR (voiced by a phonetically trained Swedish speaker, who is behind the scenes to avoid errors made by an automatic mispronunciation detection), and were given in writing in a sub-title window. Each trial proceeded as follows. When the window background changed to green, the target word was displayed and the participant was prompted to speak. Participants could ask ARTUR for repetition of the word in a normal or slow speed in order to see the difference between their own production and the correct articulations, and listen to their previous attempt. Participants were given feedback after each trial. This could be positive for a correct pronunciation, corrective the first time a participant mispronounces a word, augmented instructions for repeated errors, vague if the articulatory cause of the mispronunciation could not be determined, encouraging to get the participant to re-attempt, or giving no additional instructions. Any corrective feedback was accompanied by an animation showing the articulation for the target phonemes /r/ and /ɧ/. The results showed that audio-visual articulatory instructions were beneficial, and that participants improved their pronunciation by following the articulatory instructions indicated by the virtual teacher. However the usefulness of the vision of the tongue was not specifically evaluated.

Although VTHs containing detailed articulatory models may not be as useful for L2 learning, they do seem to improve tongue reading abilities for native speakers. Badin et al. (2010) developed a VTH that was an assemblage of individual 3D models of the jaw, tongue, lips, velum, and face of the same speaker. Magnetic Resonance

Imaging (MRI), Computer Tomography (CT) and video data were acquired from one male speaker and were aligned on a common reference coordinate system related to the skull in order to build the 3D models. Participants were tested on their identification of all French voiced oral consonants /b, d, g, v, z, ʒ, ʁ, l/ embedded in VCV context (32 VCV stimuli) where the vowels were / a, i, u, y/. They were given no feedback. To access the contribution of the tongue in relation to that of the lips and face display, participants were tested in their perception of the eight consonants in four different conditions; audio signal alone, audio signal with cutaway view along the sagittal plane without tongue present, audio signal with cutaway view along the sagittal plane with the tongue present, and the audio signal with complete synthetic face model and synthetic skin texture. The results showed that the side view presentation with the tongue yielded better consonant identification than the other presentations, indicating that some but not all articulatory information aided perception.

In summary, although using high-end technology to develop several types of VTH, and virtual tutors yields some improvements in learners' speech perception and production, there is little evidence that inclusion of detailed articulatory information, either through written instructions or through detailed animations of articulators within the VTH, affects perception or production (e.g., Massaro and Light, 2003; and Massaro et al., 2008). One reason why these models have not been as successful as one might expect, is that the learners may find it difficult to access appropriate information to help them in learning novel pronunciations from such detailed models. First, the learner likely will not have a very detailed understanding of how the different articulators (lip, jaw and tongue) contribute to the production of each individual sound. Further, the models often show a number of articulators interacting, meaning that, there are many features competing for the learner's attention. It is reasonable to assume that without explicit training, naïve learners may not know which articulator (e.g., tongue, lip or jaw) or which particular aspect of an articulator (e.g., lip-rounding, tongue retractions) is most important in effecting improvement in pronunciation. Other aspects of the modules, such as transparent skin, whilst making the model more naturalistic, may also make it more difficult for naïve learners to see the different articulators, again, making it more difficult for them to extract the appropriate information.

Another reason why learners may not have benefitted as much from this style of training is feedback. Although studies using a VTH often give feedback, this is usually given by the virtual teacher. Consequently, such feedback is usually preprogrammed and thus cannot be responsive to learner's individual difficulties.

VTH also uses the virtual teacher as shown in particular dimension to give feedback on the mispronunciation of specific phonemes. This means that if the learner was presented with the side-view of the VTH they are shown the correct tongue position (i.e., they continue to see the side-view) but if they were presented with the front-normal view of the talking head they see lips, and jaw movement, but not the three articulators together. This means that the learner does not see how the articulators combine to produce a particular phoneme.

As a result, the present study takes a different approach to production-based training for second language learning. The production training in this study combines basic articulatory phonetics training with face-to-face teaching based on computer-based animations for the training of English vowels. The animations are presented via a custom-made computer interface, CALVin (Computer Assisted Learning for Vowels interface) and are based on schematic mid-sagittal section diagrams of the principal articulators involved in English vowel production (i.e., tongue, lips and jaw). Learners see the animation at normal speed and hear the vowel produced in isolation by a native speaker. The animation is then broken down into a series of still images that detail the sequence of articulatory movements. These still images are accompanied by written text that direct the learner's attention to the critical feature in non-technical language (e.g., tongue position, lip position and jaw movement). Additionally, a phonetically-trained instructor guided learners through each training session (cf. Hattori, 2009) and was thus able to respond to individual queries and difficulties. Learners are also able to hear and repeat native-speaker recordings of the vowels in isolation and in CVC words besides they can record themselves, playback their own recordings and compare their recordings to that of the native speaker’s.

In order to assess the efficacy of this training approach vis-à-vis more traditional approaches, a large-scale training study was conducted in which this

training method was compared with two other approaches: perception-based training (HVPT), and a hybrid training programme (HTP) that combined production (CALVin) and perception (HVPT) training.

A group of forty-six Arabic speakers took part in five training sessions and were assigned randomly to one of the training conditions; 16 participants participated in 5 one-to-one production training (PT) sessions, 15 participated in perceptual training (HVPT), and 15 participated in the hybrid training programme (HTP) The hybrid condition consisted of one session of production training followed by four sessions of perceptual training (HVPT). To assess potential changes in speech production and perception, all participants completed a battery of tests before and after training. Four tasks were used to assess perception; vowel identification bVt, forced-choice, minimal pairs), category discrimination, and speech recognition in noise. To assess production, participants made recordings of the 14 English vowels bVt words they had identified in the vowel identification task, and also recorded 10 IEEE sentences.

In addition, this study also aimed to shed further light on the nature of the link between production and perception, in particular whether speech perception and production share common underlying mechanisms. If production and perception share common representations then it would seem likely that training in one domain would lead to improvements in the other. That is, training in production should lead to improvements in both production and perception, and training in perception should lead to improvements in both production and perception. In this scenario, both HVPT and CALVin-based production training should lead to similar amounts of improvement in production and perception. However, it is not clear that training in one domain leads to improvement in another (see p. 82-83 for discussion). In this case then, production training may lead to improvement in production but not perception, and perception training may lead to improvement in perception but not production. Only a training programme that incorporates both production and perception training (HTP) would lead to improvements in both domains.

Chapter 5 Investigating the domain-specificity of phonetic training: A comparison of different phonetic training methods for vowel perception and production in Arabic learners of English.

5.1 Methodology

5.1.1 Participants

A total of 57 Arabic participants were tested. Eleven participants did not complete the training sessions. Of these, 2 scored over 90% in the pre-test vowel identification task and so were considered too advanced (see Iverson & Evans, 2009), and 9 did not show up after the first session. This gave a total of 46 (18 male) participants who completed the training and all pre- and post-tests. Participants were randomly assigned to one of the three training types: Production Training (PT, 16 participants: 10 HP, 6 LP), High-Variable Phonetic Training (HVPT, 15 participants: 9 HP, 6 LP) and Hybrid Training combining both production and HVPT training (HT: 15 participants: 8 HP, 7 LP).

All participants were residents in London at the time of testing. They were mainly 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. The participants were 18-39 years old (median 27 years old). They had begun to learn English when they were 5-35 years old (median 13 years old). The participants had 3- 69 months experience of living in an English speaking country (median 4 years). However, almost all participants (4 out of 46 participants) informally reported more daily interactions with speakers from their home country, or with none-native English speakers of other language backgrounds. All participants reported no history of speech or hearing problems.

Participants were recruited to have a range of abilities with English. This was to increase individual variation in vowel perception and production accuracy within one training group rather than solely focusing on between-group variation. Although it is common in the literature to control for experience, the current study aims to take

advantage of the individual variability in order to better understand the acquisition of L2 vowels, rather than treat it as a confound that should be removed (see Iverson & Evans, 2007). As pointed out by Iverson and Evans (2007), experience is only one of the factors that may determine whether individuals are good or poor at acquiring L2 phonemes; motivation, aptitude, and the type of experience, are also important. Thus, rather than just examining between-group differences, the analysis also addressed individual differences in perception and production (e.g., the relationship between vowel identification and vowel production).

In order to evaluate their English language skills, independently of their abilities in speech perception and production, all participants completed the written grammar section of the Oxford placement test (Allan, 1992). The 3 different training groups were very similar in terms of their ability on this task. Scores ranged from 21-49 out of 50 for the PT group, median 29.5; 19-46 out of 50 for the HVPT group, median 30; and 19-36 out of 50 for the HT group, median 29. These scores were used in the analysis in order to investigate any potential effects of ability with English on speech perception and production. Participants were classified as either High Proficiency (HP) or Low Proficiency (LP) based on the overall median score: those who scored 29.5 or above were classified as HP and those who scored below 29.5 were classified as LP. This resulted in the following distribution across training conditions: PT - 10 HP, 6 LP, HVPT - 9 HP, 6 LP, and HT - 8 HP, 7 LP.

In addition, 10 Standard Southern British English speakers (4 males) participated in the study. They were 18-40 years old (median 21 years old), recruited from the UCL Psychology pool, and all were from the south of England. These participants rated Arabic learners' production for accent and intelligibility, and recorded the same /b/-V-/t/ words recorded and identified by Arabic learners to give normative data.

5.1.2 Apparatus

The pre-and post-tests were conducted in a quiet room with stimuli played over headphones (Sennheiser 555) at a user-controlled comfortable level. Stimuli were played via a Dell Inspiron N5040 laptop with digital output built-in audio sound card. The same PC laptop was used to collect responses via an experimental interface.

Recordings were made using a digital audio recorder (Zoom H2 Handy Recorder, digital stereo or 4-channel audio option) at 44,100 kHz, 16-bit resolution.

All perceptual training (HVPT and HPT training conditions) was completed by participants in their own time. Participants in these conditions all used the UCL Vowel Trainer, which applies HVPT method with different speakers, using minimal pairs, and giving direct feedback on learners' responses (Iverson & Evans, 2009). Participants provided their own laptops which they brought to the first training session, and the training software was installed onto this machine. 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 (PT and HPT training conditions) was completed with an instructor (the author) with the aid of a custom-made computer programme, CALVin (Computer-Assisted Learning for Vowels interface, see section 5.2.4 for details). Each session took 40 minutes and took place in quiet rooms using the laptop, the headphones, and the Zoom H2 Handy recorder.

Stimuli for UCL Vowel Trainer, pre- and post-test vowel identification, and category discrimination tasks were the same as those used in Iverson & Evans (2009) and Iverson et al., (2012) (section 5.1.3.2). These stimuli were recorded in an anechoic chamber at UCL with 44,100 Hz 16-bit samples per second, and later band-pass filtered (60-20000 Hz with a smoothing factor of 10) and downsampled to 22050 Hz. Stimuli for the sentence recognition in noise task were taken from existing recordings made at UCL in a sound-attenuated booth. Stimuli for the PT (CALVin) were recorded in a sound-attenuated booth at UCL with 44,100 Hz 16-bit resolution, by a native southern British English speaker.

5.1.3 Training stimuli

5.1.3.1 PT (Production Training) & CALVin design

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 (heed /iː/, hid /ɪ/, head /e/, heard /ɜː/, hard /ɑ/, hoard /ɔː/, had, /æ/, hud /ʌ/, hod /ɒ/, who'd /uː/) and four diphthongs (how'd, /aʊ/, hoed /əʊ/, hayed /eɪ/, hide /aɪ/). These were then grouped into five clusters; High/front: /iː/, /ɪ/, /e/ (e.g., heed, hid, head); Open: /æ/, /ʌ/, /ɒ/ (e.g., had, hud, hod); Central/low back: /ɜː/, /ɑː/, /ɔː/ (e.g., heard, hard, hoard); Back: /uː/, /aʊ/, /əʊ/ (e.g., who'd, how'd, hoed); and Diphthongs: /eɪ/, /aɪ/ (e.g., hayed, hide). These clusters were expected to be highly confusable for Arabic learners of English, based on hierarchical cluster/Euclidian distance analysis on previous English vowel identification by different group of Arabic learners of English, (see chapter 3).

The speaker recorded two example words for each vowel, giving a total of 28 examples (2 examples for 14 vowels), and an example of each vowel in isolation. Example words had a CVC, CCVC or CVCC structure; *back, bad, barn, park, bed, Ben, bird, burn, shout, blouse, caught, forks, feet, heat, fight, white, hate, fate, shoot, flute, cod, cost, code, cone, hit, fit, bud, bun*. The words were selected to be familiar to L2 learners, and as far as possible, were orthographically unambiguous. To ensure that the isolated vowels were as naturalistic as possible, the speaker recorded each isolated vowel after the keyword. While recording the words and the isolated vowels the speaker could see a word on the screen, and was instructed to produce the word and isolated vowels with a falling intonation contour. In order to make the isolated vowels longer than they might be produced within words whilst maintaining the distinction between tense and lax vowels, the speaker was instructed to utter the word first, then a longer version of the vowel on its own. The speaker recorded 3 repetitions of each word and the isolated vowel. The best recording was used for the stimuli.

All stimuli were band-pass filtered (60-20000 Hz with a smoothing factor of 10) and then saved into individual wav files so that they could be embedded within the training software.

The CALVin interface was designed using a Graphical User Interface (GUI). CALVin was designed to be used as a training/teaching tool similar to any interface software, to support the acquisition of English vowels. The interface enables learners to listen to isolated vowels, view animations of the isolated vowels, contrast recordings of different vowels within and/ or between pre-defined vowel clusters, and to record their own voice so that they could compare their own production of a given vowel with that of a native speaker (see Fig. 5.1). Within CALVin, vowels are grouped into 5 different clusters and within each cluster there are interactive buttons that allow participants to access the different functions.

Users can switch between clusters by clicking on the cluster buttons (Fig. 5.1). Keywords for each vowel within the cluster can be heard by pressing the keyword buttons (Fig 5.1). These serve as a substitute for using IPA transcription. Additionally, each keyword button enables the user to access the other functions, i.e., the animation and step-by-step instructions for that vowel.

The animations show the movement of the articulators when producing the isolated vowel. Each animation consists of 12 images that start from the neutral position of the articulators and end at the same neutral position. The vowel target (i.e., the position of the articulators at the midpoint of production of the vowel) was the base for creating the images that shape the animation, and the other 11 images were gradual movement from the neutral position of the articulators moving towards this, and then gradually moving back to the neutral position. The vowel target was based on existing descriptions of vowel production (e.g., Ladefoged, 1996).

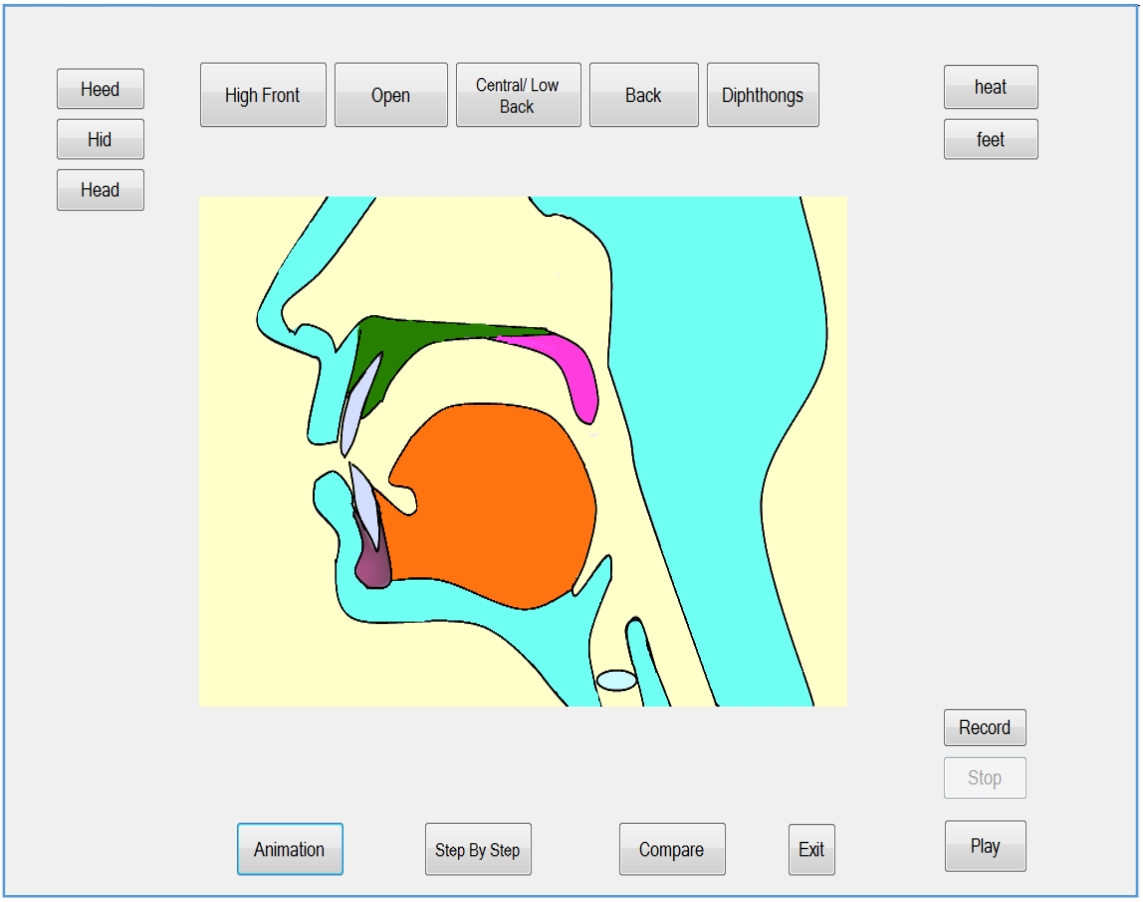


Figure 5.1: A snapshot of CALVin software showing the animated mid-sagittal section CALVin in a neutral position in the centre of the screen, the keywords on the top-left, the clusters on the top-middle, and the example words on the top-right. The animation, step-by-step instructions and compare buttons are on the bottom, and the record and play-back/stop button on the bottom right.

The animations were intended to be accurate approximations rather than faithful physiological animations; there is much between-speaker variability in vowel production and so it was felt that these idealized animations would be as beneficial for learners, if not more so. The images were created using paint.NET (a graphics editor program), which enables the instructor to view a series of images at once and edit the shape of the moving articulators, while controlling the size of the images so that they can be used later for animation. In order to ensure that the sequence of the created images could be used in the animation, the 12 images were first gathered as different layers and animated using GIMP2 (a raster graphics editor program). Where the transition between images was not smooth, images were edited again using paint.NET (Paint.Net 4.0.3), and then checked by re-animating the images in GIMP2.

Once the animation was judged to be as naturalistic as possible (i.e., smooth), the individual images for each animation were gathered with the wav file of the corresponding isolated vowel recording using GIF (animation maker program). The length of the animation was adjusted to fit with the duration of the wav file of an isolated vowel. For example, if a vowel had a duration of 1.106 s., and there were 12 images, then the duration was divided by the number of images that shaped the animation of the vowel, (i.e., 1.106 divided by 12, meaning that each image was displayed for 0.09 s).

Users are also able to view the changes in configuration of the articulators for a vowel using the "step-by-step" button. This function gives access to a still-image of the target configuration with the three moving articulators, the tongue, jaw and lips, highlighted in successive steps. Each step has written instructions on how to position the articulators (see Fig. 5.2), with a bold highlight on the tongue in one picture, jaw on the other, and on the lips in the third picture. There are arrow buttons that allow the learner to navigate forwards or backwards through the sequence.

Clicking on the "compare" button displays the still images of the vowel target for each vowel within a cluster along with the correspondence keyword for the vowel (see Fig. 5.3) along with a 'back' button that allows participants to go back to the main window of CALVin.

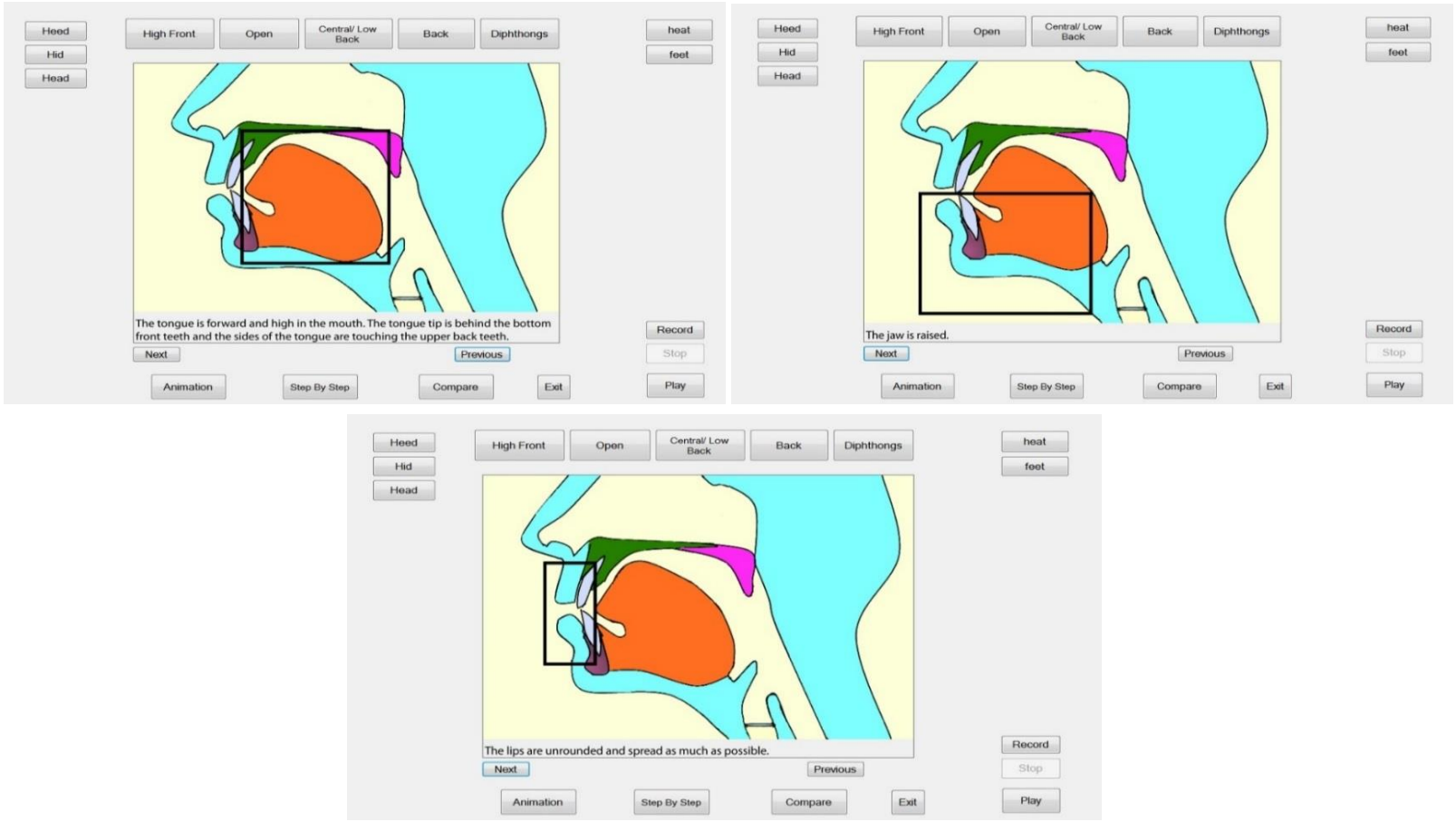


Figure 5.2: An example of the step-by-step button display for the/i/ vowel. The first picture highlights the tongue, the second the jaw and the third the lip movements. All pictures have written instructions underneath, with 'next' &'previous' buttons to allow learners to navigate between pictures.

When the user has clicked on a vowel within a cluster, they are able to access two example words which they can click on and listen to. These are not accompanied by animations. Users are also able to record, stop, and play-back their own productions (Fig 5.1). Participants can record and replay their own production so that they can compare their own production with that of the native speaker's, as well as getting feedback from the instructor. The play-back button was added because it has been argued that "self-perception" (i.e., listening to one's own production) helps in learning L2 sounds (cf. Baker & Trofimovich, 2006). Successive recordings are not stored; once a user records another sound, the previous recording is automatically erased.

The software is used along with a small mirror that allows the participants to see their jaw and lips and compare their production of the different vowels with the aid of the still images (Compare button: Fig. 5.3) and feedback from the instructor.

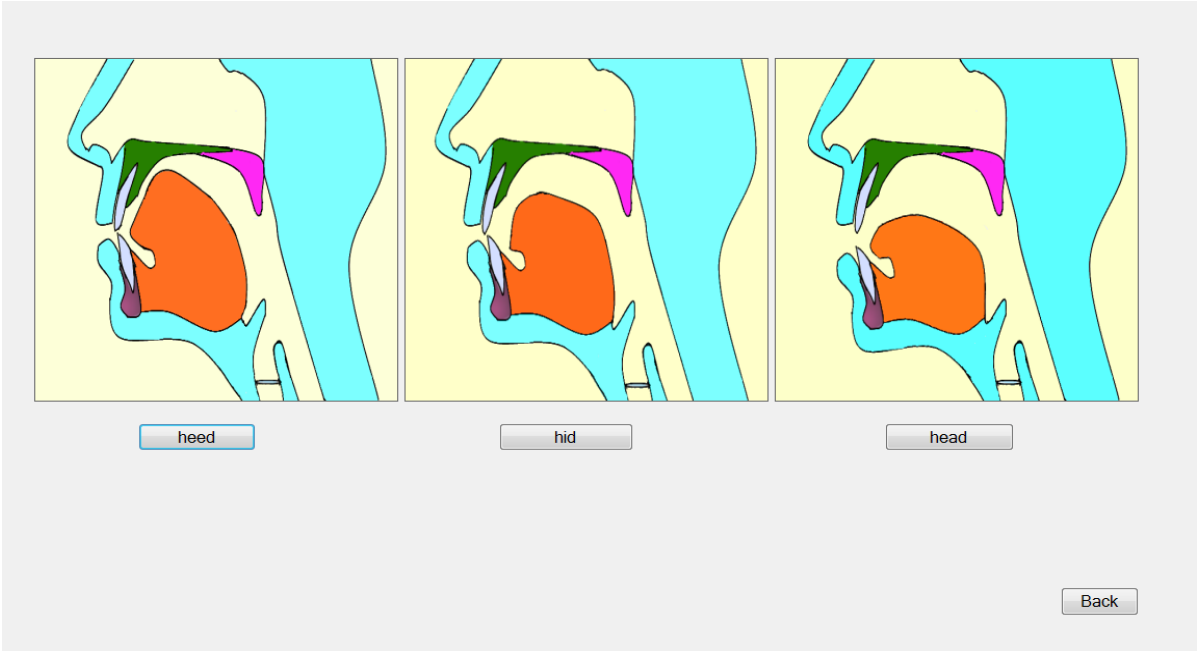


Figure 5.3: A snapshot of the compare button. Learners can use this to see the difference in the configuration of the articulators for each vowel target, and can click on the keyword to listen and repeat the vowel.

5.1.3.2 Perceptual training using the UCL Vowel Trainer

Training in the perceptual training condition was conducted using the UCL vowel trainer (Iverson and Evans, 2009). This trainer adapted the HVPT technique

(Logan et al., 1991). The training stimuli were the same as in Iverson and Evans (2009). The recordings of the training words were made by five speakers of British English, three female and two male. The vowels were divided into four clusters: /e/, /ɑː/, /æ/, /ʌ/ (e.g., *pet, part, pat,* and *putt*); /iː/, /ɪ/, /aɪ/, /eɪ/ (e.g., *feel, fill, file, fail*); /ɒ/, /əʊ/, /ɔː/ (e.g., *was, woes, wars*); and /uː/, /aʊ/, /ɜː/ (e.g., *shoot, shout,* and *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). The first three clusters comprised vowels that were mutually confusable, and the last cluster (/uː/, /aʊ/, /ɜː/) was formed of the vowels that were not as strongly clustered with others. There were 10 sets of minimal pairs for each of these clusters, giving a total of 140 words. Each speaker recorded each word twice and each recording was used during training. During the recording, words were displayed individually in a random order to avoid list intonation effects.

5.1.3.3 Hybrid Training

The HTP method consisted of combination of the production and HVPT training methods. As such, the stimuli in this condition were the same as those in the production and HVPT training conditions.

5.1.4 Stimuli for pre- and post-tests

5.1.4.1 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/ words made by 10 speakers of British English (5 male, 5 female), all from the south of England. None of these words and speakers were used in the training corpus, such that all pre- and post-tests measured generalization to new stimuli. The speakers read the /b/-V-/t/ words: *beat* /iː/, *bit* /ɪ /, *bet* /e/, *Bert* /ɜː/, *bat* /æ/, *Bart* /ɑː/, *bot* /ɒ/, *but* /ʌ/, *bought* /ɔː/, *boot* /uː/, *bait* /eɪ/, *bite* /aɪ/, *bout* /aʊ/, and *boat* /əʊ/. English vowels that would create non-words in the /b/-V-/t/ context (e.g., /ʊ/) were not included in the study.

5.1.4.2 Speech recognition in noise

The stimuli were recordings of the phonetically balanced IEEE Harvard sentences (Rothauser et al., 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 sentence lists were recorded by a male SSBE speaker. The stimuli for the SSBE speaker were taken from existing recordings at University College London. All the recordings were made in 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 (Sennheiser HD 555) in a quiet room.

5.1.4.3 Production

Participants recorded the same 14 English /b/-V-/t /words that they were asked to identify in the vowel identification task, and 10 IEEE sentences, specifically the first 10 sentences (i.e., the first block).

5.1.5 Training Procedure

5.1.5.1 Production training

There were five sessions of training, each conducted with an instructor. During the course of the 5 sessions, all 14 English vowels (10 monophthongs, 4 diphthongs) were trained. Participants completed no more than one session per day, and the entire 5 sessions were completed over 1-2 weeks. Each session lasted no more than 40 minutes. Additionally, participants completed a practice session lasting 10 minutes before starting the first session. In this session, the instructor familiarised the participants with the software, and explained the relationship between the different positions of their tongue, jaw and lips and the resulting vowel sound. Participants were asked to produce back, front, open, and closed vowels (e.g., *heed*, *had*, *who'd*) and the position of the articulators was explained to them in each with the help of a hand mirror so that participants could see their lip and jaw movements. Every effort was made to avoid technical language.

Participants were familiarized with the training software, and were then asked to produce back, front, open and closed vowels while looking at a mirror so that they could get sense of the various tongue, lip and jaw positions.

Each session followed broadly the same structure. At the beginning of each session, participants were trained on all 5 clusters for 10 minutes, where they spent more time on the vowel/contrast they found most difficult. Each session started from the high/front cluster and ended with the diphthongs cluster. Then they spent 20 minutes training on one cluster, one each in a fixed order from high/front to the diphthong clusters. The remaining 10 minutes of training reviewed the trained cluster in the context of 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. All training was completed in English.

Training on the individual clusters proceeded as follows. For each vowel, participants were instructed to start off by clicking on a keyword within a cluster to hear the vowel. By doing so, they were made aware that if they clicked on a keyword the corresponding examples and isolated vowel changed to be that of the vowel in the keyword. For instance, participants were guided to the target vowel cluster (e.g., high-front vowels), and then clicked on one of the keywords (e.g., *heed*) to hear the version of the vowel. They were then guided through the articulatory process involved in producing the vowel using the animation function in CALVin (Fig. 5.1). They viewed the animation and were then guided to the step-by-step function that described the principal articulatory positions of the tongue, jaw and lips (see Fig. 5.2). For example, for the vowel /iː/ they saw a still-image of the vowel target for /iː/, highlighting the position of the tongue in one image, the jaw in another, and the position of the lips on a third image. Each image was accompanied by a written description of how to position the articulators. After viewing the step-by-step instructions, participants practised producing the vowels. First they produced the isolated vowel, the key word and then finally the example words. They were then asked to record themselves producing the

isolated vowel, keyword, and the example words, play back their recordings, and compare them with the native speaker's production.

During the training session, 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. For example, when participants confused the /ɪ/- /e/ contrast, the instructor explained that there is a slight drop in the jaw from /ɪ/ to /e/, and the tongue is lowered when producing /e/, whilst when producing /ɪ/ the jaw is closer. After explaining the difference, the instructor asked the participant to produce the contrast (i.e., /ɪ/- /e/) while looking at their production in a hand-mirror, guiding them to focus on the difference in jaw position.

5.1.5.2 Perceptual training using the UCL vowel trainer

There were 5 sessions of HVPT along with an initial 14-trial session. 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).

On each trial, participants heard a stimulus word. Then they saw three or four minimal pair alternatives, and clicked on the one that they thought they had heard. For instance, participants heard the word *fox* and then chose from three response options; *folks*, *fox*, or *forks*. The stimulus word was played before the response options were shown, with the intent that their initial recognition of the word would be open set (e.g., not primed by the response alternatives), even though they were presented with a closed-set response (Iverson & Evans, 2009). In case the subject was not familiar with the response word, each word response was accompanied by a higher frequency or common alternative that had the same vowel (e.g., *go*, *pot*, *born*). These example words were the same whenever that vowel appeared as a response.

Participants received feedback on their responses. If participants clicked on a correct response, they saw "Yes!" on the computer screen accompanied by a cash register sound, then heard the word one more time. If participants clicked on the wrong response, they saw "Wrong" on the computer screen accompanied by two tones with

descending pitch, heard the correct response played, then heard a four-stimulus alternating series of the correct word and the incorrect response. For example, if the stimulus word was *folks* and they clicked on the *forks* response, they would hear an alternating series of *folks*, *forks*, *folks*, and *forks*. This was intended to help them learn the distinction between these two words.

As described in Iverson & Evans (2009), each training session was made up of 225 trials. The first 70 trials were 5 repetitions of the 14 vowels in a random order, the next 85 trials were chosen adaptively based on the participant's errors, and the last 70 trials were also 5 repetitions of the 14 vowels in a random order. This design ensured that all participants were trained on all the 14 vowels at the beginning and end of each session, while allowing for some training to be customized to fit the needs of each individual subject. The trials that were chosen adaptively were selected randomly, with the selection probability of an individual vowel being weighted by combining the proportions of misses and false alarms of the vowel. That is, the probability of the vowel being selected increased when it was identified incorrectly, or when that vowel was chosen incorrectly as a response when another stimulus had been played (Iverson & Evans, 2009).

The stimulus words on each trial were chosen randomly for each vowel. That is, if the trial was intended to have an /i/ stimulus, the computer programme randomly choose one of the ten minimal-pair stimulus words that had this vowel. This random selection was blocked, such that each of the ten minimal-pair word sets was used once before the list was recycled.

5.1.5.3 Hybrid Training

As for other training programmes, the HTP programme consisted of five training sessions; one session of PT (CALVin) that took approximately 40 minutes, preceded by a practice session of 10 minutes, and 4 sessions of HVPT (UCL Vowel Trainer).

The PT session followed broadly the same procedure as described above (PT procedure). Participants spent 10 minutes at the beginning and end of the session on

all vowels, though they were encouraged to spend more time with the vowels that they found more difficult. The middle part of the session lasted 20 minutes and focused on each cluster in turn. The order started from the high/front cluster to the diphthongs cluster, giving around 4-5 minutes training on each cluster.

After finishing the articulatory session, the UCL vowel trainer (Iverson& Evans, 2009) was installed on the participant's laptop. Participants followed the same training procedure as those in the perceptual training condition, but were asked to finish four rather than five sessions of the training.

5.1.6 Procedure for pre- and post-tests

5.1.6.1 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 selected on each trial (i.e., all 10 were mixed within the same block) and were randomly mixed to make the task more equivalent to the category discrimination task described below, which requires having stimuli from different talkers.

5.1.6.2 Category discrimination

This task was the same as that described in Iverson et al., (2012). Participants heard three English /b/-V-/t/ words on each trial, which were 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., they completed 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. For example, participants heard 11 pairs of /ɪ/-/e/ words, with each pair played six times. Within each pair, 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 (see chapter 3). 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.

5.1.6.3 Speech recognition in noise

The participants performed a sentence recognition task in which they listened to IEEE sentences (Rothauser, et al., 1969) in noise. They were asked to verbally repeat what they had heard, with the experimenter logging 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 sentence was presented only once. 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). The procedure began with an easy stimulus with an SNR of +10dB (i.e., above threshold) which enabled participants to tune in to the speaker. After each correct response, the level of SNR decreased in 8 dB steps (i.e., became harder), until the first reversal (i.e., an incorrect response). The SNR then changed in 2dB steps for a further eight reversals after the first reversal.

A one-up/one-down procedure was used, and when participants repeated all five keywords aloud to the researcher, the sentence was scored as correct. The sentence was scored as incorrect when participants repeated only two, one or none of the keywords. The SNR remained the same when the participants repeated two keywords and this did not count as a reversal. The procedure therefore converged on a 50% identification level. The test terminated when participants had completed eight reversals or after 20 stimuli were presented.

5.1.6.4 Production

English Vowel Production. All participants recorded the /b/-V-/t/ words 3 times. The recordings of all participants in the pre- and post-test were analysed acoustically, and were also given to 10 Standard Southern British English (SSBE) listeners for identification judgments, following the same procedure as in the vowel identification described above.

For the acoustic analysis, only the monophthongs were analysed. The clearest two repetitions (i.e., no hesitation, lip-smack, good voice quality) 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 the physiological differences, whilst preserving other sources of variation in the acoustic measurements (Adank et al., 2004).

IEEE sentences. Participants were asked to record the first ten sentences of the IEEE sentences list (i.e., list 1) at a normal reading pace. After testing had been completed, one sentence was selected to be used for accent ratings; *'Glue the sheet to dark the blue background*'. This sentence was chosen as it was judged to be accent revealing. The sentence contains various features that Arabic learners of English typically find challenging, e.g., /uː/ which is often produced more like Arabic /o/, /ɑː/ ,and a middle word consonant cluster, 'back-ground'-/kgr/, that Arabic speakers find hard to produce.

Ten SSBE listeners rated tokens from this training study as well as those from participants who completed PT in Saudi Arabia (see chapter 6). They rated a total of 220 tokens; 2 sentences (1 pre, 1 post) x 46 participants (16 HVPT, 15 HT, 15 PT, and 9 PT in Saudi) x 2 repetitions. Stimuli were presented in a random order over four

blocks. The rating session took a maximum of 90 minutes, with a short break in-between each block. Listeners were asked to rate the speech of the speaker on a Likert scale from 1 (strongly-accented) to 7 (native-like accent). The raters were encouraged to use the whole of the scale. Listeners were unaware that they would hear each speaker more than once, and that these speakers had completed any training. Additionally, they were not given any information about their first language background.

5.2 Results

For the following results, when the mixed-effects models were chosen, they were chosen using a top-down approach, in which ineffective factors were excluded after all possible factors had been included.

5.2.1 Vowel Identification

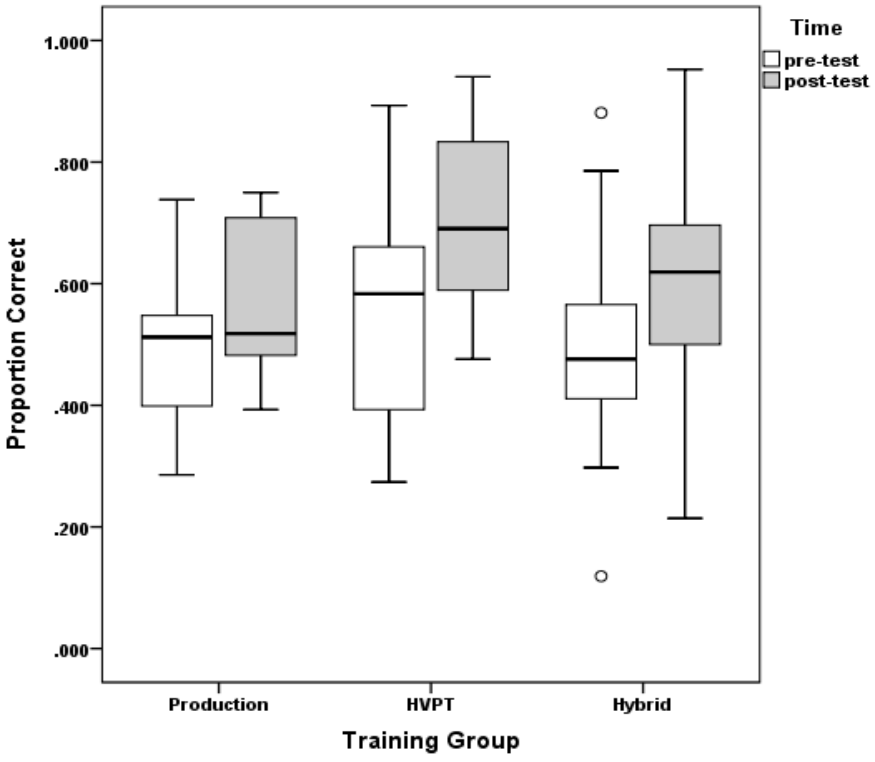


Figure 5.4: Boxplots showing the proportion correct of vowel identification scores at the pre- and post-tests across training groups

Fig. 5.4 displays the vowel identification accuracy for Arabic learners of English in the three different training types; PT, HVPT, and HTP. The proportion correct of the vowel identification task appears to improve from pre- to post-test in the HTP and HVPT groups, but not much in the PT group. However, when split by proficiency, there appears to be a small change in performance in the LP group in the PT condition (see Fig. 5.5).

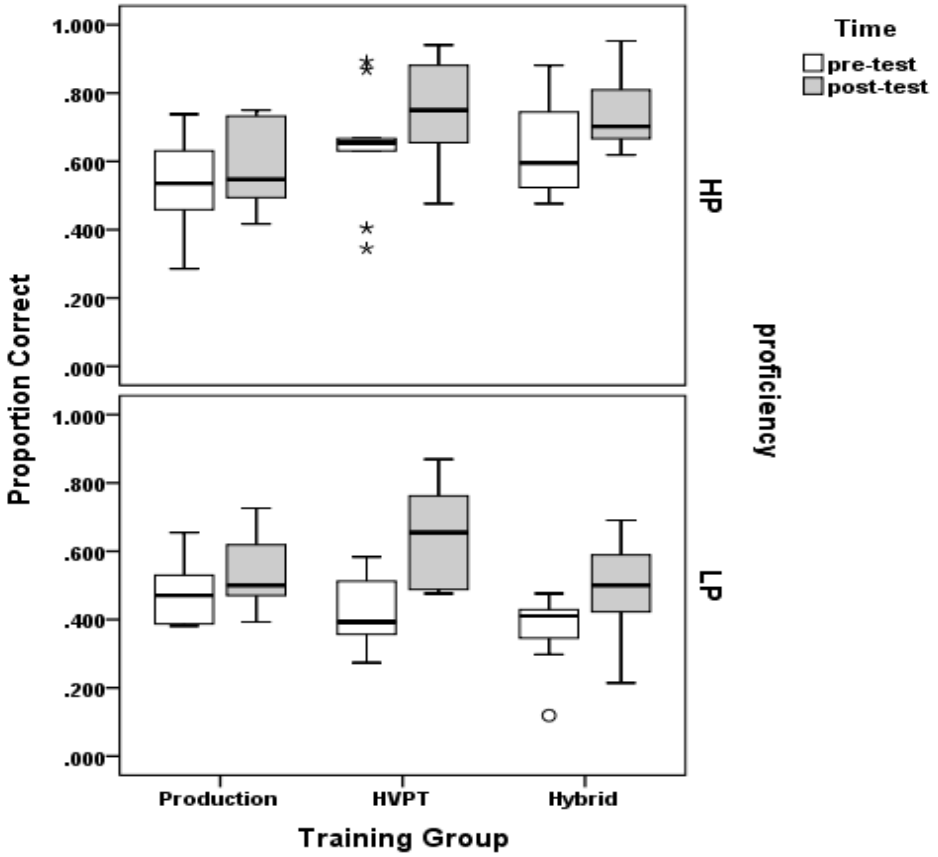


Figure 5.5: Boxplots showing the proportion correct of vowel identification scores at the pre- (white boxes) and post-tests (grey boxes) across training groups (PT, HVPT, and HT), split by proficiency level (HP= High proficiency, and LP= low proficiency).

In order to verify the effect of training and proficiency on vowel identification improvement, a logistic mixed effects model was built for the binomial identification responses (i.e., correct/incorrect). The best fitting model to the data was fit by the Laplace approximation with time (pre- and post), proficiency (HP, or LP), and training group (type of training) coded as fixed factors, and participant and stimulus coded as random factors. The best model excluded the three-way interaction between group,

time and proficiency which suggests no significant effect of this three-way interaction. The random factors were added with random slopes for time (pre/post-tests), so that the difference in the pre- and post-tests could be calculated per stimulus, and per participant in a crossed-design. This is because all participants listened to the same set of stimuli. Although the stimuli were produced by different speakers, nesting the stimulus into the speaker was not the best fit to the data (i.e., even if the stimuli had been produced by different speakers, and speaker had been added in a nested design to the stimuli, it had no significant effect).

The logistic regression model showed that the main effect of training group was not significant χ^2 (2) = 1.888, p>.05, indicating that training type did not affect vowel identification performance differently. That is, everyone improved regardless of training type. The main effect of time (pre-post) was highly significant χ^2 (1) = 35.685, p<.0001, indicating a change from pre- to post-tests. The orthogonal planned contrasts confirmed a change from pre- to post-test, b=-0.3112, SE=0.05082, z=-6.125. p<.0001; participants improved in their vowel identification scores from pre- to post-test. There was a significant effect of proficiency χ^2 (1) = 5.406, p<.05, which suggests that participants with different proficiency levels were affected differently by the training. The orthogonal planned contrasts showed that the LP learners improved more in their vowel identification accuracy after training than the HP group in all training conditions, b= 0.25251, SE=0.08948, z=2.822, p<.05. Although there was no significant effect of training group, there was a significant interaction between group and time χ^2 (2) = 13.78, p<.05, demonstrating that some groups improved more from pre- to post-test than others. The orthogonal planned contrasts showed that the HVPT yielded significantly more improvement in vowel identification accuracy than did the PT from pre- to post-test, b= -0.2123, SE= 0.05734, z= -3.704, p<.0001. This suggests that HVPT is more effective in improving identification accuracy than the PT programme. However, the orthogonal planned contrasts showed no significant difference between learners' performance in HVPT and the HT programme. There was no significant difference between participants'' performance in the production and the HT program.

There was also a significant two-way interaction between group and proficiency, χ^2 (2) = 7.819, p<.05. The orthogonal planned contrasts showed a significant difference between HP and LP participants in the HTP group compared with HP and LP participants in the two other training groups (i.e., PT and HVPT). That is, in the HVPT and the PT groups, LP but not HP learners improved in their vowel identification accuracy, but both LP and HP participants in the HTP programme improved, b= -0.17220, SE= 0.06758, z= -2.548, p<.05. Even after removing the outliers in the HP group in the HVPT, there was no significant change in HTP learners after training.

5.2.2 Category discrimination

Fig. 5.6 shows the category discrimination scores for each word-pair at the pre- and post-test for each training group. Overall there does not appear to be much change from pre- to post-test in discrimination performance. In order to test this observation, a linear mixed model was built for the category discrimination data. The best-fitting model included time (pre-post), as a fixed factor, and participant and word pair coded as random factors. Interestingly, the main effect of time was significant, χ^2 (1) = 27.99, p<.001, which suggests that there was a change from pre- to post-test. The orthogonal planned contrasts showed a significant difference from pre- to post-test, b=- -0.045, SE= 0.00862, pMCMC<.001 indicating that at least for some word pairs, discrimination improved. Visual inspection of the boxplots indicates that this effect would be most likely driven by improvements in discrimination of *bit-bet, bart-bot, bart-but* and *bat-but*. The best model excluded training group, and the interaction between time and proficiency, indicating no significant effect of these factors.

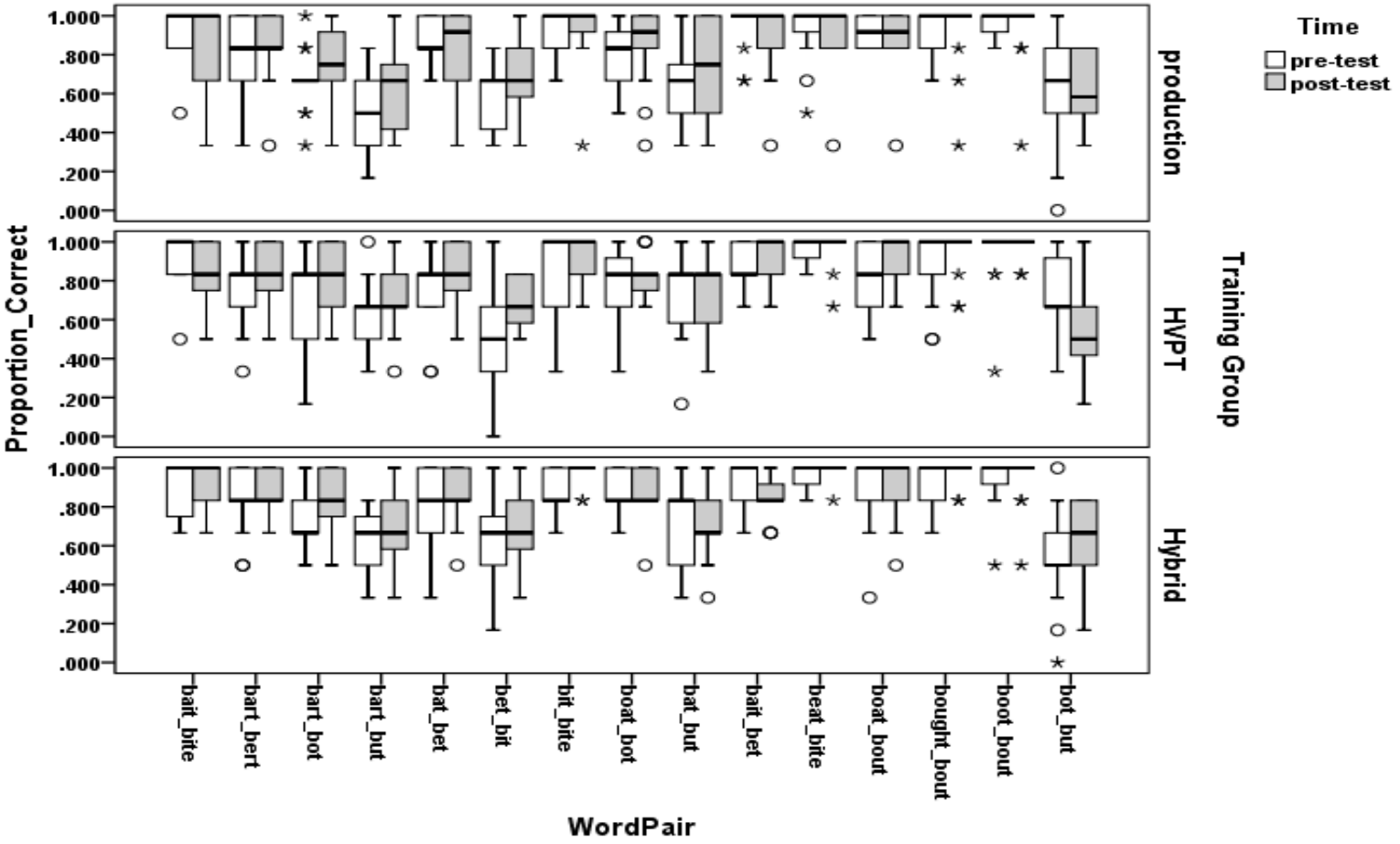


Figure 5.6: Boxplots showing category discrimination accuracy (proportion correct) for each word-pair across training types (PT, HVPT, and HT) at the pre- (white boxes) and post-tests (grey boxes).

5.2.3 Speech recognition in noise

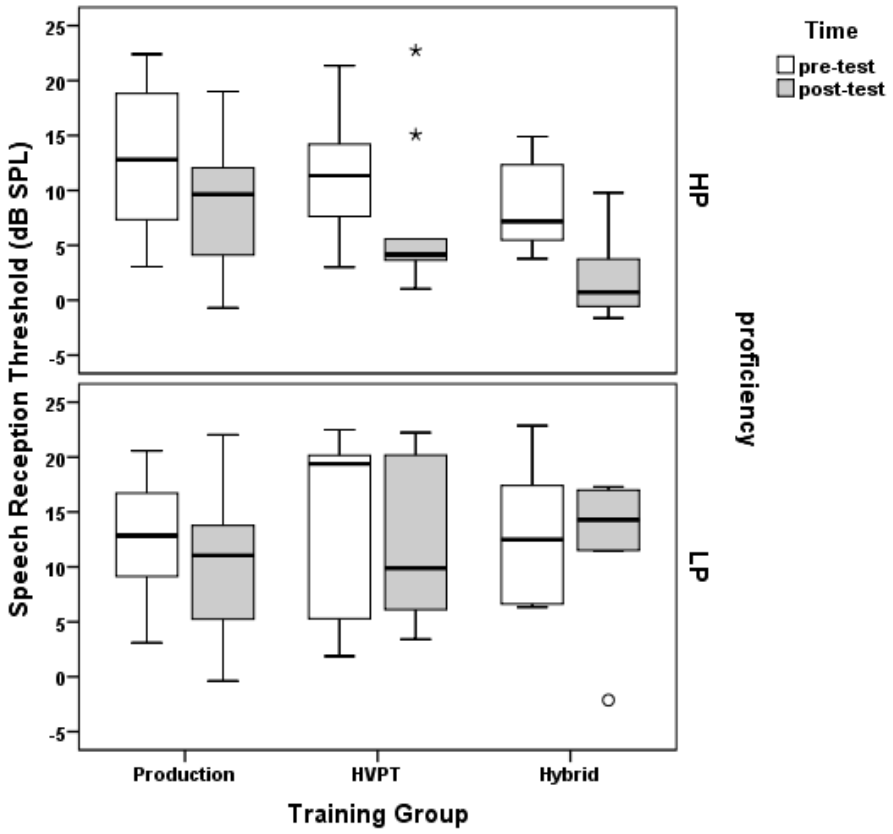


Figure 5.7: Boxplots of speech perception threshold for L2 listeners across training groups at the pre- (white boxes) and post-tests (grey boxes) and split by proficiency level; High Proficiency (HP: top panel), and Low Proficiency (LP: bottom panel).

As displayed in Fig. 5.7, HP learners appeared to improve more in performance on the speech in noise task after training than did LP learners. Additionally, there appeared to be an effect of training; HP learners who completed HTP and HVPT, appeared to improve more than those who completed PT. In order to test these observations, a linear mixed model was built to examine any potential changes in the speech ratio threshold (SRT) scores. The best fitting-model for the data included time (pre-post), and proficiency (HP, LP) coded as fixed factors, and participant as a random factor. The main effect of training group was not significant $p>.05$. However, the main effect of time was significant, $\chi^2(1) = 17.48$, $p<.001$, indicating that learners improved from pre- to post-test. The planned contrasts showed a significant

improvement in speech recognition in noise from pre- to post-test, b= 1.651, SE= 0.417, pMCMC<.05. The main effect of proficiency was significant χ^2 (1) = 5.708, p<.05, confirming that listeners with different proficiency levels performed differently. The orthogonal planned contrasts confirmed that the HP learners improved more in their speech recognition in noise than did LP learners, b= -1.997, SE= 0.835, pMCMC<.05. This supports the observation from the boxplot (Fig. 5.7) that the HP listeners improved more in speech recognition in noise after training.

5.2.4 Speech production

5.2.4.1 Acoustic Analysis of /b/-V-/t/ words

In order to avoid multiple comparisons, the monophthongs were divided into three groups; Group 1: *beat, bit, bet, bert*, Group 2: *bat, but, bart*, and Group 3: *boot, bought, bot*. The analysis of F1& F2 for each vowel group is presented first, followed by an analysis of duration, again, for each vowel group. To enable comparison of male and female talkers, formant frequency measurements were normalized using Lobanov's method (Lobanov, 1971).

5.2.4.1.1 Spectral analysis

Group 1: Beat, Bit, Bet, Bert. As shown in Fig 5.8, there is some evidence of change in F1 from pre- to post-test, but little change in F2 values. In order to look for any spectral changes for the vowels after training, separate linear mixed models were built for F1, and F2.

The best fitting-model for F1 included training group (PT, HVPT, HTP), time (pre-post), and proficiency (HP, LP), which were coded as fixed factors, and excluded the interactions between time and proficiency and time and group, indicating that these were not significant.

The best-fitting model also included participant and stimulus, coded as random factors, with a random slope for time. The main effect of the training group was approaching significance, χ^2 (2) = 5.956, p=.05, which may suggest that type of training affected changes in F1 values from pre- to post-test. The orthogonal planned

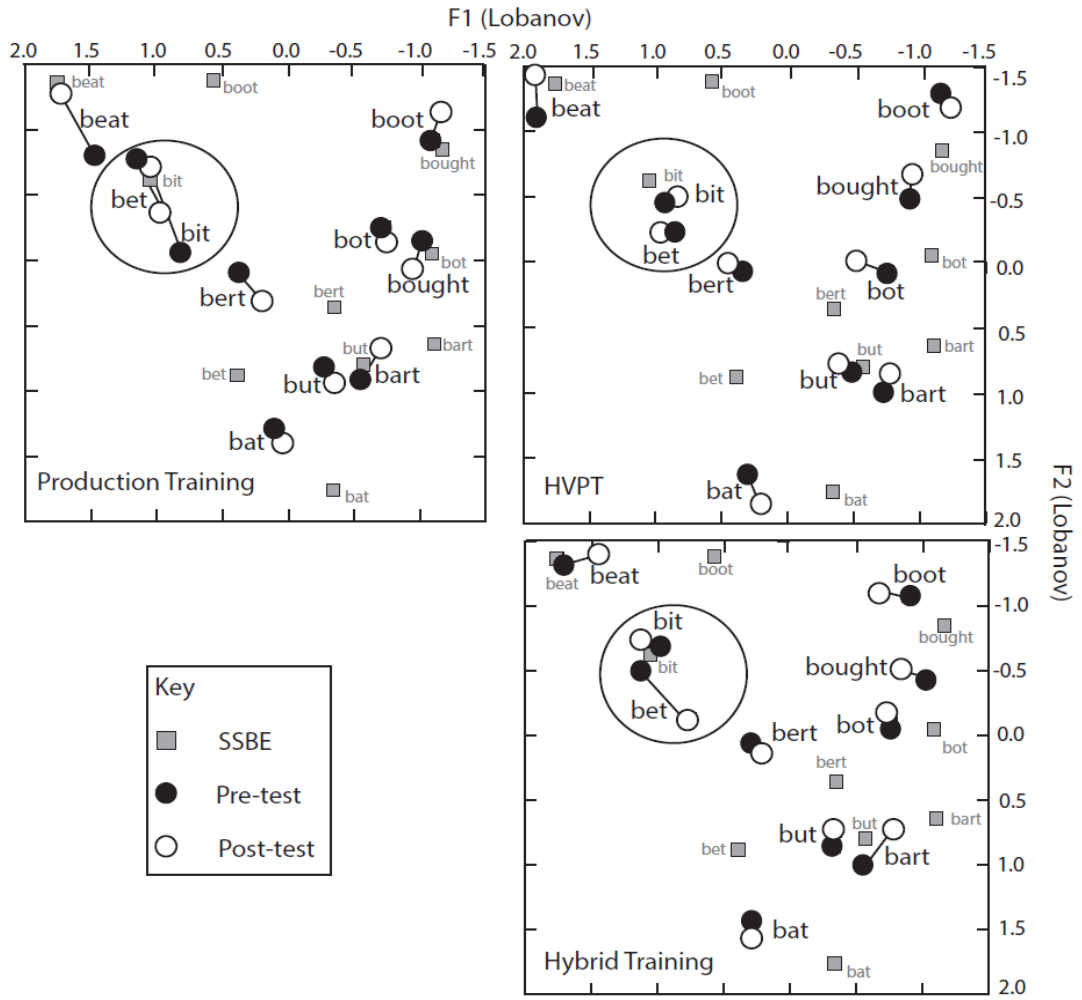


Figure 5.8: vowel plots for L2 speakers in the three training groups; PT, HVPT, HTP, compared to that of the SSBE speakers. The formants values were normalised using Lobanov's method

contrasts indicated that learners in the PT and HT groups changed their F1 but those in the HVPT group did not, b= -0.0161, SE= 0.0288, pMCMC<.05. Specifically, after training, learners in the PT and the HT produced the vowel /ɪ/ with a lower F1 value, and the vowel /e/ with a higher F1 value, such that the values were closer to those of native speakers (see Fig. 5.8).

However, the main effect of time was not significant p>.05, and there was no significant interaction between time and training group. Lastly, there was no main effect of proficiency, p>.05.

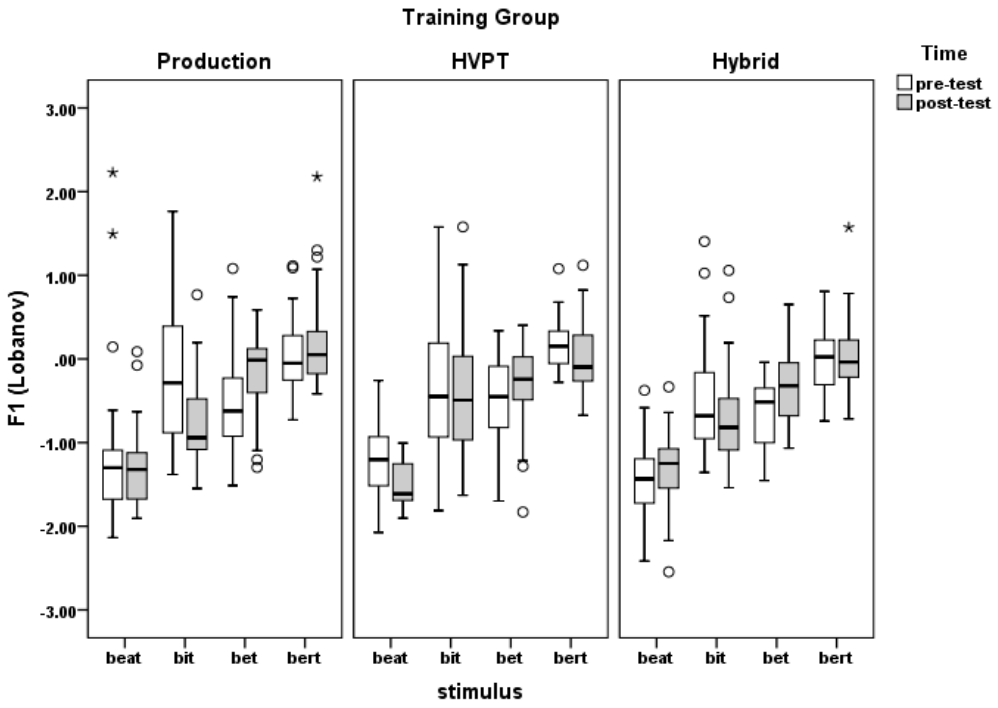


Figure 5.9: Boxplots showing F1 values for vowel group 1 (beat, bert, bet, bit) produced by learners in the three training groups at the pre- and post-test, the formants values were normalised using Lobanov's method. The F1 values for stimuli was the average of 2 repetition of a word for each speaker.

The best fitting-model for F2 included training group (production, HVPT, Hybrid), and time (pre-post) coded as fixed factors, and stimulus with a random slope as a random factor. There was no significant effect of time or training group, indicating no significant change in F2 values from pre- to post-test for any of training groups.

Group 2: Bat, But, Bart. As displayed in Fig. 5.8, there did not seem to be not much change in F1, but a slight change in the F2. In order to look for any spectral changes for this vowel group after training, separate linear mixed models were built for F1and F2.

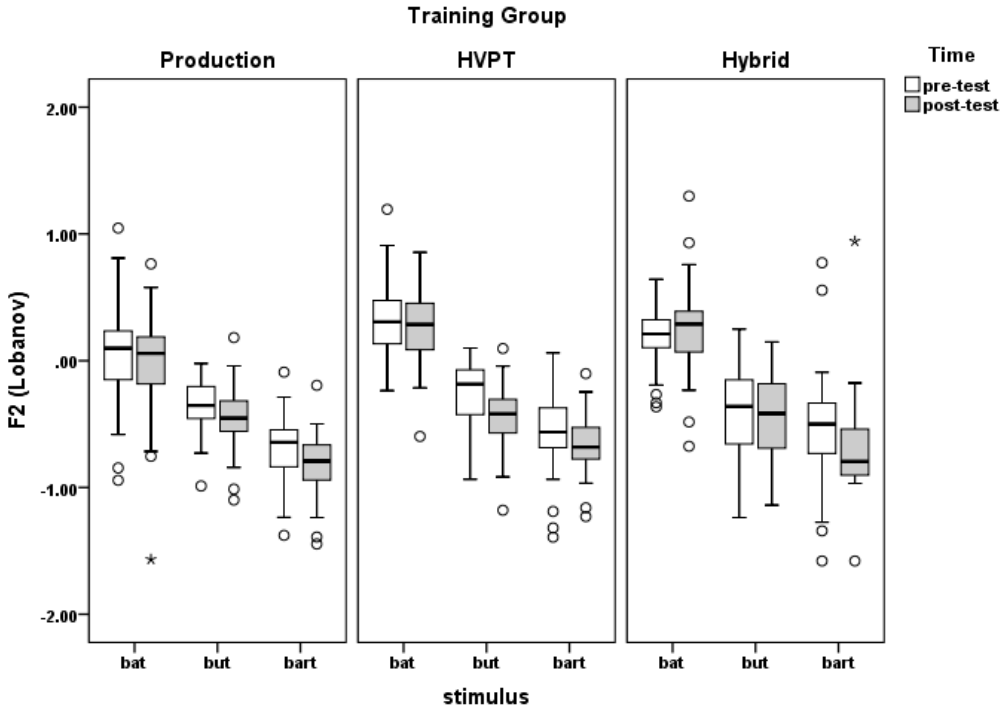


Figure 5.10: Boxplots for F2 values in pre and post-tests for vowel group 2 (bat, but, bart) produced by learners in the three training groups, the formants values were normalised using Lobanov's method. The F2 values for stimuli was the average of 2 repetition of a word for each speaker.

The best fitting-model for F1 included time (pre-post) as a fixed factor, and stimulus as a random factor. There was no significant effect of either factor indicating no significant change in the F1 value from pre- to post-test for any of training groups.

The best model for F2 included time (pre-post), and group (PT, HVPT, HTP) coded as fixed factors, excluding the interaction between time and group, and proficiency. The random factors included stimulus and participant. The main effect of time was significant χ^2 (1) = 10.069, p<.05, suggesting a change in F2 values from pre- to post-test. The orthogonal planned contrasts showed a significant change in F2 values from pre- to post-test, b= 0.0399, SE=0.0125, pMCMC<.05.

The main effect of group was significant χ^2 (2) = 7.5499, p<.05, indicating that learners in different training groups used different F2 values. The planned contrasts showed a significant change in F2 values in the vowels produced by speakers in PT compared to those in the HVPT, b=0.078, SE=0.0299, pMCMC<.05 (see Fig. 5.10).

PT speakers change their production such that it was closer to the native speakers' F2 values for these vowels (see Fig. 5.8). There was no significant difference in F2 values between the PT and the HTP groups, or between HTP and HVPT groups.

Group 3: Bot, Bought, Boot. As displayed in Fig. 5.8 there appear to be some small changes for F1, but not for F2 values. In order to look for any spectral changes after training, separate linear mixed models were built for F1, and F2. The best fitting-model for F1 included time (pre-post) and group (PT, HVPT, HT) as fixed factors and stimulus as a random factor. There was no significant effect of any factors suggesting that there was no significant change in F1 values. Likewise, the best fitting-model for F2 included time (pre-post) and group (production, HVPT, hybrid) as a fixed factors and stimulus as a random factor, and also did not show any significant effects.

Summary. In brief, these results suggest that there were spectral changes in vowel production but that these were limited to a small number of vowels. Specifically, the changes were in the F1 values for /ɪ/ and /e/ and F2 values for /ʌ/ and /ɑ/; in both cases, learners adjusted their formant frequency values to better match those of native speakers (see Fig. 5.8). Additionally, these changes were limited to those learners who completed the PT and the HTP programmes (i.e., where the training included explicit training in speech production).

5.2.4.2 Duration

Group 1: Beat, Bit, Bet, Bert. As displayed in Fig. 5.11 there appears to be some change in the duration values for long vowels (i.e., *bert, beat*) from the pre- to post-test. In order to verify the effect of different training types on vowel duration, a linear regression mixed effects model was built for the duration data using the duration of the vowels (beat, bit, bet, bert) in milliseconds (continuous scale).

The best-fitting model included training group (PT, HVPT, HTP), proficiency (HP, LP), and time (pre-post) coded as fixed factors, and participant and stimulus coded as random factors. The random factors were added with random slopes for pre/post testing, so that the difference in the pre- and post-tests could be calculated per stimulus for each participant in a crossed-design. This was done because, all

participants produced the same set of stimuli. A three-way interaction between time, training group, and proficiency was excluded by the model indicating that this interaction was not significant.

The main effect of time was significant, χ^2 (1) = 6.774, p<.05, indicating that participants' vowel duration changed from pre- to post-test. The orthogonal planned contrasts showed a significant change in vowel duration from pre- to post-test, b=-8.403, SE= 3.446, pMCMC<.05=.003. Fig. 5.11 indicates that this is because learners changed their productions to be longer in duration that than of native speakers. In particular, they produced long vowels (*beat*, *bert*) with longer duration.

The main effect of the training group was also significant, χ^2 (2) = 16.39, p<.001 suggesting that different training programmes yielded different changes in production. The planned contrasts showed a significant change in the vowel duration produced by HTP participants compared to HVPT and PT participants, b= 6.3884, SE = 2.4543, pMCMC<.05.

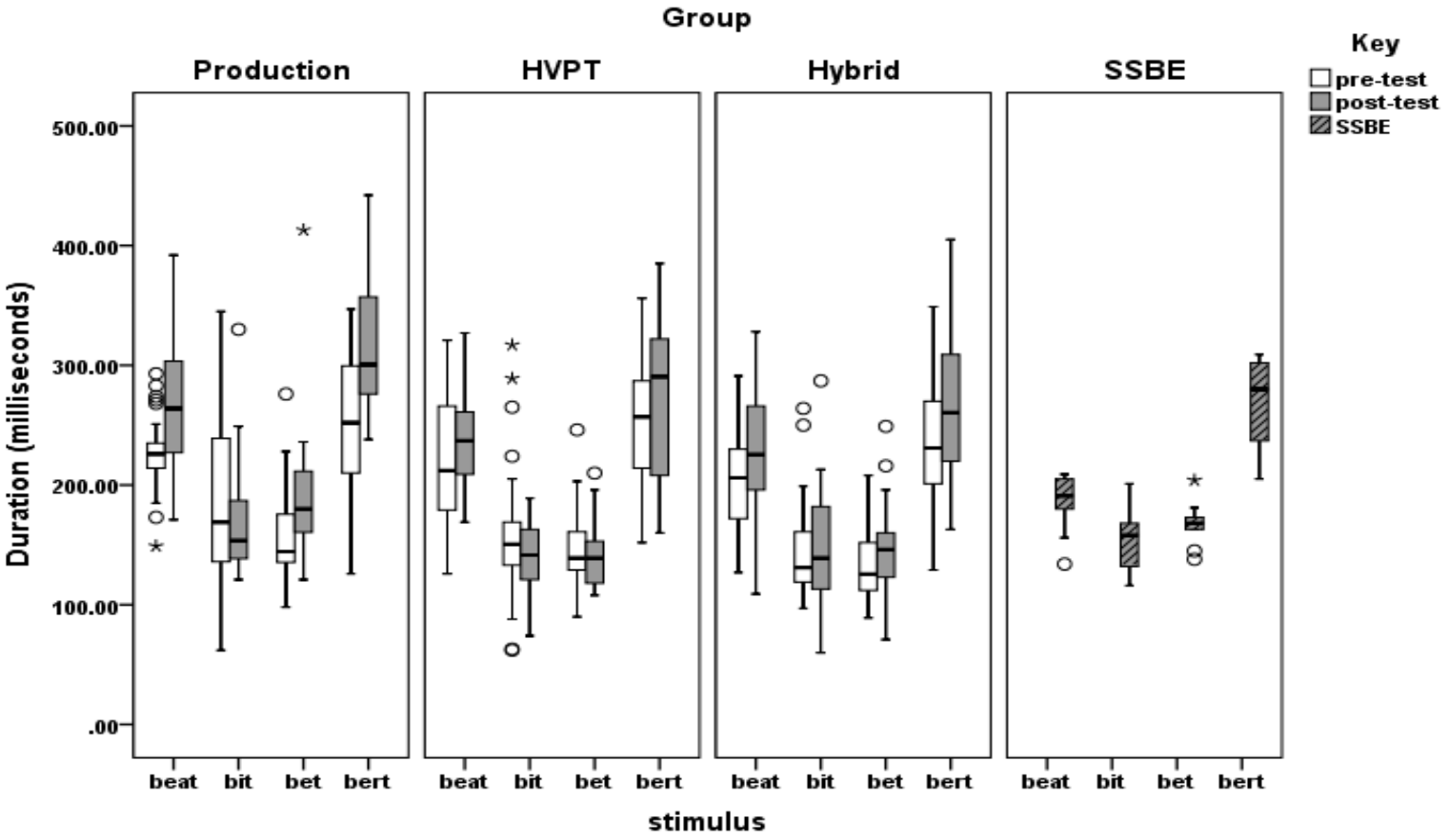


Figure 5.11: Boxplots of vowel duration in milliseconds for vowel group 1 (beat, bit, bet, bert) at the pre- (white boxes) and post-tests (grey boxes) across training groups (PT, HVPT, HTP), compared to the SSBE speaker group (dark grey with lines).

Additionally, the planned contrasts also showed that PT participants changed more than HVPT participants, b= -12.8415, SE= 4.27389, pMCMC<.001. However, the interaction between time and group was not significant, p>.05 which may suggests that though some may have changed more than others, all participants changed the duration of these vowels to some extent as a result of training, (see Fig. 5.11). The main effect of proficiency was not significant, p>.05.

Group 2: Bat, But, Bart. As displayed in Fig. 5.12, there were some changes in vowel duration from pre- to post-test, with potentially more change in the HP group than the LP group (see Fig. 5.13). In order to investigate any potential changes in duration for these vowels, a linear regression model was built for duration of the vowels (*bat, but, bart*) in milliseconds (continuous scale).

The best fitting-model included training group (PT, HVPT, HTP), proficiency (HP, LP), and time (pre-post) coded as fixed factors, and participant and stimulus coded as random factors. The random factors were added with random slopes for pre/post testing (as above).

The main effect of time was significant, χ^2 (1) = 34.3831, p<.0001, suggesting that learners changed the way in which they produce these vowels from pre-to post test. The orthogonal planned contrasts showed a significant vowel duration from pre- to post-test, b=-10.1478, SE=1.7306, pMCMC<.001, such that after training, learners produced these vowels with a similar pattern to native speakers (see Fig. 5.12). The main effect of training group was significant χ^2 (2) = 7.8842, p<.05, which indicates that learners in different training groups behaved differently. The orthogonal planned contrasts showed a significant difference in improvement in vowel duration for participants in the HVPT group compared to the PT group, b=-11.4053, SE=6.0273, pMCMC<.05, and a significant difference in improvement between participants in the HTP group and those in the other two training groups (HVPT & PT) b= 7.1020, SE=3.4612, pMCMC<.05. Inspection of the data revealed that this was because change in duration for these vowels from pre- to post-test was greater in the HTP group than in the HVPT and PT groups, where there were some changes in duration (see Fig. 5.12).

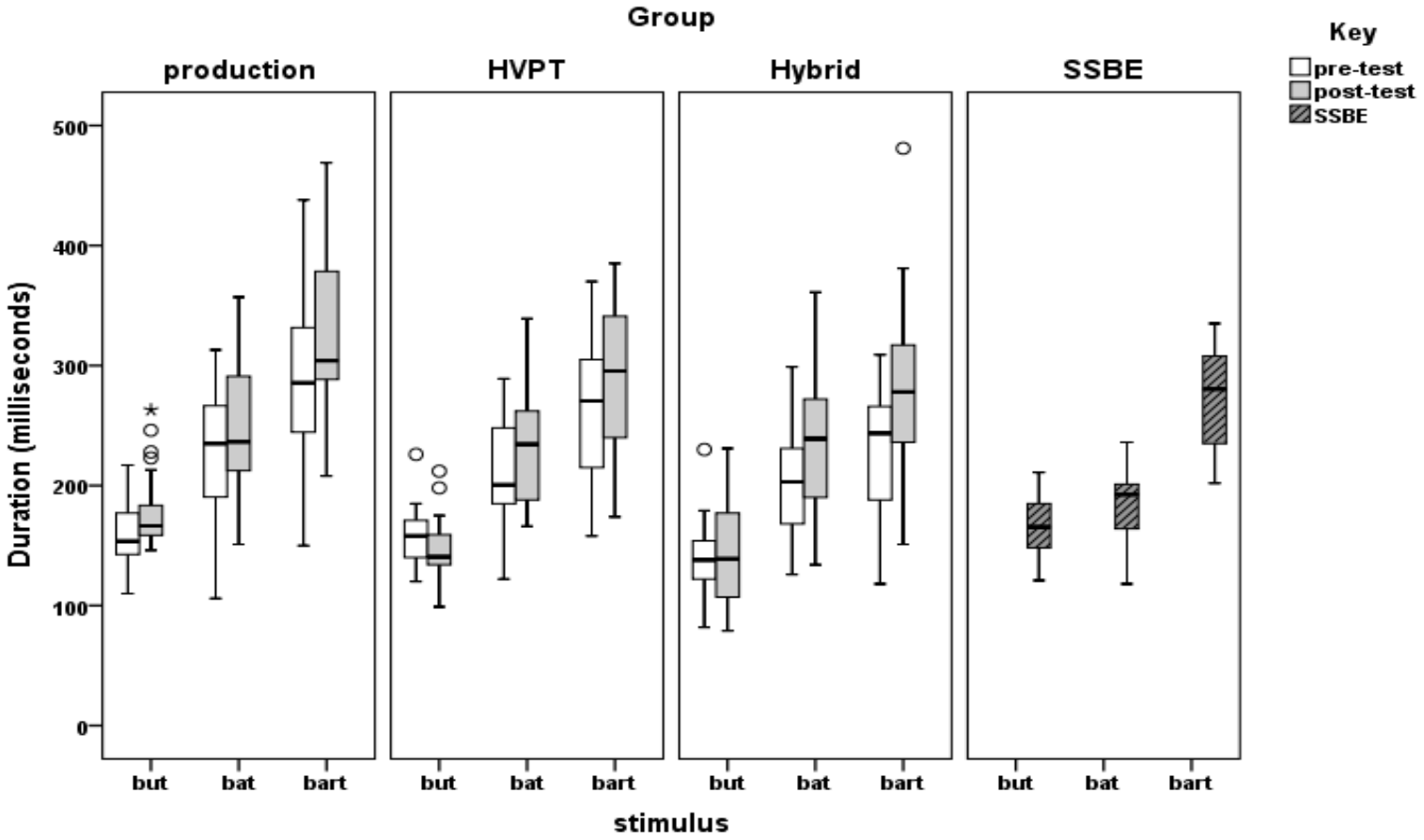


Figure 5.12: Boxplots showing vowel duration in milliseconds for vowel group 2 (bat, but, bart) produced by L2 learners at the pre-test (white boxes) and post-tests (grey boxes) for the three training groups compared to that of the SSBE speakers (dark grey boxes with

Although there was no significant effect of proficiency, the interaction between time and proficiency was significant, χ^2 (1) = 5.8506, p<.05, suggesting that participants with different proficiency levels changed their vowel duration differently from pre- to post-test. The orthogonal planned contrasts showed that the HP participants behaved differently from the LP participants, b= -4.1860, SE= 1.7306, pMCMC<.05; the HP group produced the *bart* vowel with longer duration at the post test compared to the LP group (Fig 5.13), which is longer than that of the native speakers .

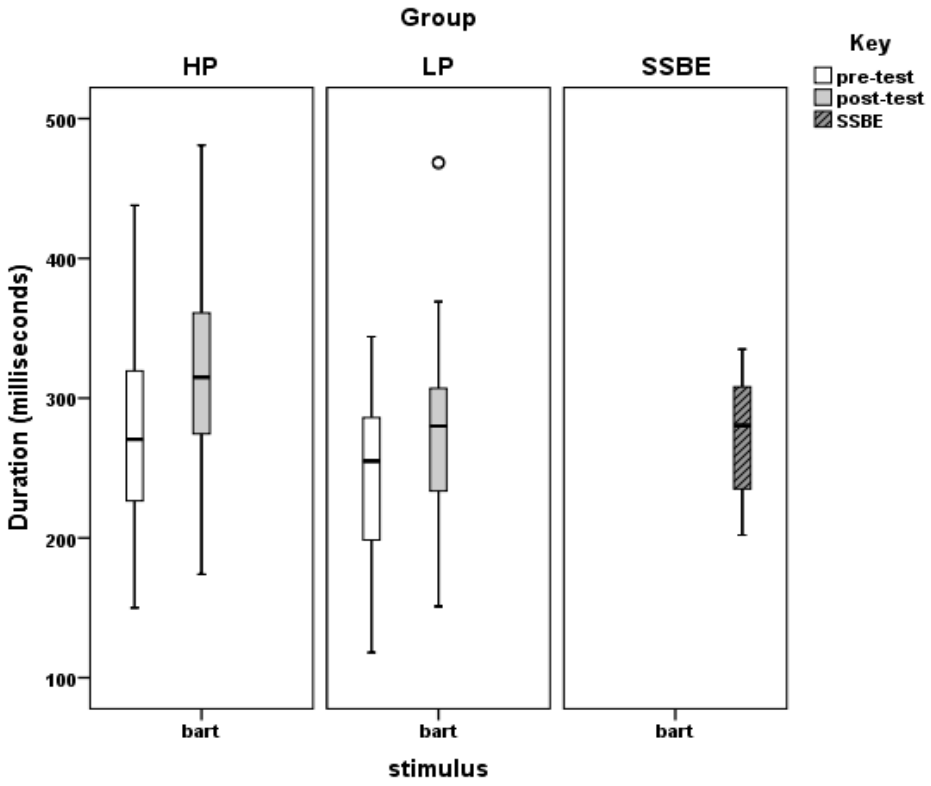


Figure 5.13: Boxplots showing vowel duration for vowel group 2 (bat, but, bart) at the pre and post-tests for the three training groups (PT, HVPT, HT), split by proficiency level (HP, LP), compared to that of the SSBE speakers.

Group 3: Bot, Bought, Boot. As displayed in Fig 5.14, participants in the PT and HTP groups made some changes to their production of these vowels, in particular, the vowel (bought). In order to verify the effect of different training types on vowel duration, linear mixed models were built for the duration data based on the duration of the vowels (boot, bot, and bought) in milliseconds (continuous scale). The best fitting-model included training group (PT, HVPT, HTP), proficiency (HP, LP), and time (pre-post) coded as fixed factors, and participant and stimulus coded as random factors.

The random factors were added with random slopes for pre/post testing as explained above. The main effect of time was significant, χ^2 (1) = 9.210, p<.001, indicating a change in the vowel duration from pre- to post-test. The orthogonal planned contrasts indicated a significant change in vowel duration from pre- to post-test, b= -12.7787, SE= 4.2106, pMCMC<.001; vowels tended to be produced with longer duration after training.

The main effect of group was significant χ^2 (2) = 19.4937, p<.001, suggesting that the vowel duration values were different across training groups. The orthogonal planned contrasts indicated a significant change in vowel duration from pre- to post-test for participants in the PT group compared to those of participants in the HVPT group, b= -13.52, SE= 5.170, pMCMC<.05. After training, participants in the PT group changed their vowel duration for the long vowels so that it was closer to that of the native speakers (see Fig. 5.19). Proficiency level was not significant p>.05, indicating that proficiency level did not affect the duration changes for these vowels.

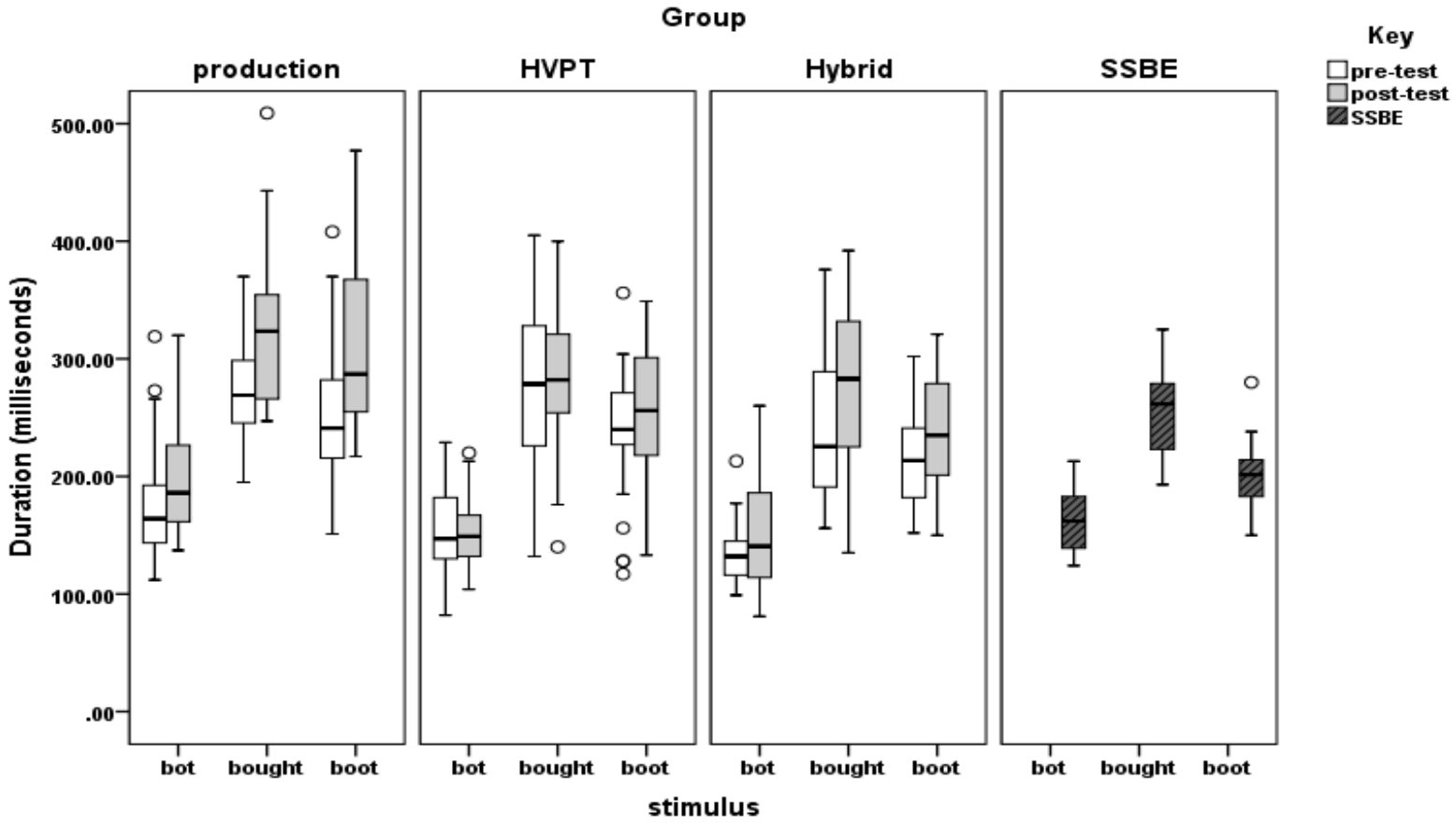


Figure 5.14: Boxplots showing the vowel duration in milliseconds for vowel group 3 (bot, boot, bought), at the pre- (white boxes) and post-test (grey boxes) for participants in the three training groups (PT, HVPT, HTP), compared to SSBE speakers (dark-grey with lines).

Summary. Briefly, the results showed some durational changes after training, specifically for individuals in the PT and the HT groups. Both groups produced the vowels (*beat, bert, bart, bought,* and *boot*) with longer duration that they did at the pre-test. Although L2 learners could change their duration after training, for some vowels, they produced close duration values to that of SSBE speakers (e.g., *bert, bart*). However, for other vowels (e.g., *beat, bought*) they produced longer vowel duration than the SSBE speakers.

5.2.4.3 Vowel intelligibility and Goodness ratings

/b/-V-/t/ recordings. As shown in Fig. 5.15, L2 learners tended to be more intelligible after training, though the amount of improvement was not large (median at pre-test=.64, and at post-test=.71, SD. pre-test=.146, SD. at post-test=.174**).**

In order to test this observation and test for any potential effects of training type on Arabic learners' intelligibility, a logistic mixed effects model was built for identification data based on the correct/incorrect binomial responses. The best fitting-model included time (pre-post), training group (PT, HVPT, and HTP), and proficiency (HP, LP) coded as fixed factors, and participant and stimulus coded as random factors. The random factors were added with random slopes for pre/post testing, so that the difference in the pre- and post-tests could be calculated for each stimulus, and for each participant in a crossed-design, as all participants listened to the same set of stimuli. The best fitting logistic mixed-effects model on identification accuracy demonstrated that there was a significant main effect of time, χ^2 (1) = 9.418, $p < .05$, indicating that there was a change in intelligibility from pre- to post-test (i.e., that learners were more intelligible after training).

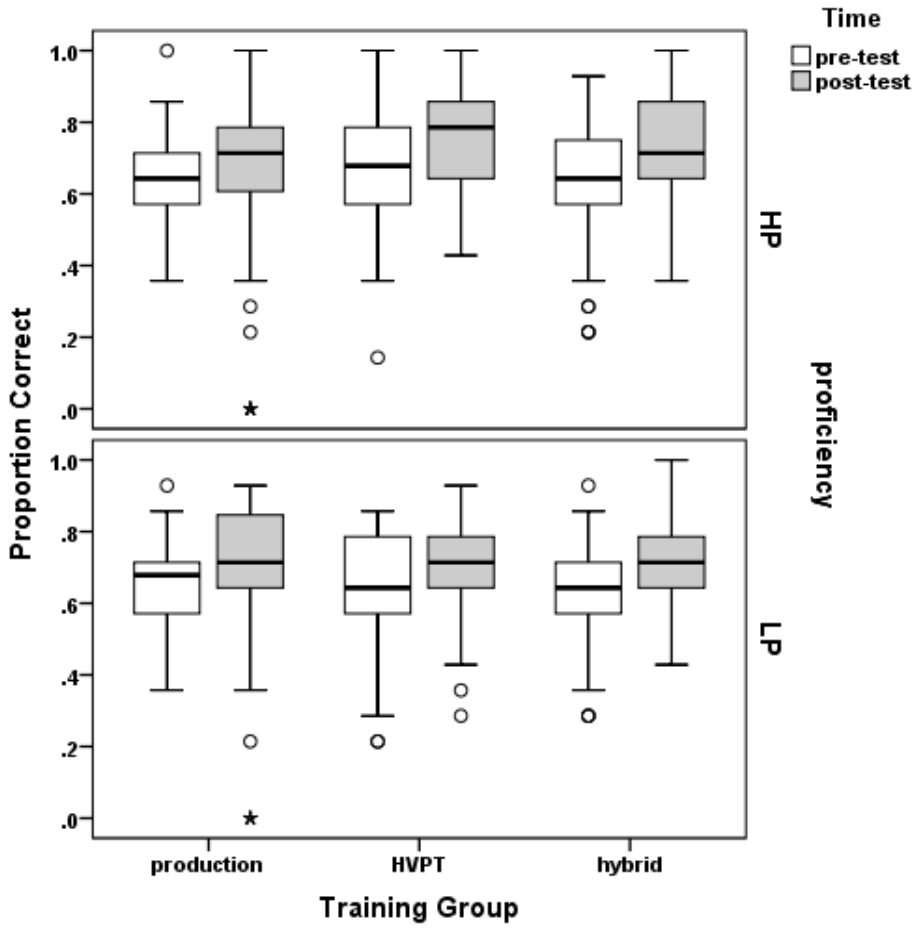


Figure 5.15: Boxplots showing vowel intelligibility scores (proportion correct) by SSBE listeners (N=10), identifying vowels produced by L2 Arabic speakers at the pre- (white boxes) and post-test (grey boxes). The top half shows the accuracy score for vowels produced by HP groups, and the bottom half for the accuracy scores for the LP groups in the three training groups (PT, HVPT, HTP).

The planned contrasts for the pre- and post-test identification verified that there was a significant change from pre- to post-test, b=-0.21766, SE=0.07037, z-3.093, p<.05. As might be expected, there was also a significant effect of proficiency, χ^2 (1) = 106.616, p < .0001. The orthogonal planned contrasts indicated that HP learners were more intelligible overall than the LP ones, b= 0.2277, SE=0.02131, z=10.684, p<.001.

There was no significant effect of training group, p>.05, suggesting that there was no difference between learners in different training groups. However, the model indicated a significant interaction between training group and proficiency level, χ^2 (2) = 52.091, p < .001, suggesting that learners with different proficiency levels were affected differently by training type. The orthogonal planned contrasts showed that

HP learners in the HVPT were more intelligible than the HP learners in the PT group, b= 0.0885, SE= 0.0293, z= 3.014, p<.05, and HP learners in the HTP group, b= -0.220, SE= 0.0307, z= -7.185, p<.001.

In order to investigate whether particular vowels were harder to identify than others, confusion matrices for pre- and post- tests were calculated (see Tables 5.1 & 5.2). Inspection of the data showed that there was improvement from pre- to post-test in particular vowels, namely; *bit, bet,* and *bought* which were not well identified at the pre-test (Table 5.1) but, with the exception of bought, improved such that they had similar intelligibility levels as other vowels at the post test (Table 5.2). The improvement also tended to be greater in the PT & HTP groups compared to the HVPT group. The amount of improvement for the PT group was 19% (bit), 21% (bet), and 26% (bought). For the HTP group, it was 16% (bit), 27 % (bet) and 9% (bought), but for the HVPT group it was 4% (bit), 13% (bet) and 19% (bought). Learners in the PT and HT groups tended to improve more in *bit* and *bet* than those in the HVPT group, though it is important to note that the HVPT group were more intelligible in their production of *bet* at the pre-test than those in either the PT or HTP groups (PT = 48%, HVPT = 69% HTP = 44%). *bought* was not well identified at either the pre- or post-test for any training group (Pre-test: PT = 9%, HVPT = 6%, HTP = 13%; Post-test, PT = 35%, HVPT = 22%, HTP = 25%), but there was some improvement in intelligibility for all groups.

	response															
		bait	**bart**	**bat**	**beat**	**bert**	**bet**	**bit**	**bite**	**boat**	**boot**	**bot**	**bought**	**bout**	**but**	**Total**
	bait	61	0	1	0	1	5	2	29	0	0	0	0	0	1	100
	bart	3	73	2	0	2	1	1	0	0	0	1	5	4	7	100
	bat	0	5	68	0	7	4	0	1	0	0	0	0	0	14	100
	beat	8	0	0	77	0	5	10	0	0	0	0	0	0	0	100
	bert	0	3	2	1	91	2	0	0	0	0	0	0	0	0	100
	bet	2	0	5	3	1	53	34	0	0	0	0	0	0	1	100
	bit	0	0	2	7	0	40	39	11	0	0	0	0	0	0	100
	bite	2	0	0	2	0	3	2	89	0	0	0	0	0	1	100
	boat	0	0	0	0	0	0	0	0	68	4	5	9	9	4	100
	boot	0	0	0	0	0	0	0	0	15	73	4	6	0	2	100
	bot	0	0	1	0	1	0	0	0	6	1	53	9	0	28	100
	bought	0	1	0	0	0	0	0	2	61	0	4	9	20	2	100
	bout	0	0	0	0	0	0	0	0	16	0	4	8	71	0	100
stimulus	**but**	0	1	10	0	0	0	0	0	0	0	7	1	0	79	100

Table 5.1: Confusion matrix showing the percent correct of the vowels identified by SSBE listeners, averaged across the three training groups at the pre-test.

		response														Total
		bait	**bart**	**bat**	**beat**	**bert**	**bet**	**bit**	**bite**	**boat**	**boot**	**bot**	**bought**	**bout**	**but**	
stimulus	**bait**	74	0	0	0	2	6	1	17	0	0	0	0	0	0	100
	bart	0	68	3	0	1	0	0	0	1	0	3	15	4	6	100
	bat	0	7	72	0	7	7	0	0	0	0	0	0	0	7	100
	beat	8	0	0	83	0	2	7	0	0	0	0	0	0	0	100
	bert	2	5	3	0	80	5	1	0	4	0	0	0	0	0	100
	bet	1	0	9	2	2	74	10	0	0	0	0	0	0	2	100
	bit	0	0	2	4	0	33	52	7	0	0	0	0	0	1	100
	bite	0	0	0	2	0	1	1	94	0	0	0	0	0	0	100
	boat	0	0	0	0	0	0	0	0	72	1	6	6	13	1	100
	boot	0	0	0	0	0	0	0	0	9	82	1	6	0	2	100
	bot	0	1	0	0	0	0	0	0	4	3	61	10	1	20	100
	bought	0	0	0	0	0	0	0	0	45	4	4	28	17	3	100
	bout	0	1	0	0	0	0	0	0	8	0	0	4	84	3	100
	but	0	1	8	0	0	0	0	0	0	0	4	1	0	85	100

Table 5.2: Confusion matrix showing the vowels identified by SSBE listeners (percent correct), averaged across the three training groups at the post-test.

	Response													
Group	Stimulus	bait	bart	bat	beat	bert	bet	bit	bite	boat	boot	bot	bought	bout	but
PT	bet	2	1	8	1	1	48	40	0	0	0	0	0	0	1
	bit	0	0	4	5	0	36	42	14	0	0	0	0	0	0
	bought	0	3	1	0	0	0	0	0	68	0	4	9	14	2
HVPT	bet	2	0	5	3	3	69	19	0	0	0	0	0	0	0
	bit	1	0	1	4	0	45	35	13	0	0	0	0	0	0
	bought	0	1	1	0	0	0	0	0	66	1	4	6	18	4
HTP	bet	1	0	3	6	1	44	43	1	0	0	0	0	0	1
	bit	0	1	2	11	1	41	39	7	0	0	0	0	0	0
	bought	0	0	0	0	0	0	0	7	49	0	3	13	27	1

Table 5.3: Confusion matrix showing the vowels (bet, bit, bought) identified by SSBE listeners (percent correct), for the three training groups at the pre-test.

	Response													
Group	Stimulus	bait	bart	bat	beat	bert	bet	bit	bite	boat	boot	bot	bought	bout	but
PT	bet	0	0	13	6	3	69	4	0	1	0	0	0	0	4
	bit	0	0	1	10	0	21	61	6	0	0	0	0	0	0
	bought	0	1	0	0	1	0	0	0	31	0	3	35	22	7
HVPT	bet	1	0	9	0	2	82	5	1	0	0	0	0	0	0
	bit	1	0	2	0	0	51	39	7	0	0	0	0	0	0
	bought	0	0	0	0	1	0	0	0	45	8	1	25	21	0
HTP	bet	1	0	5	0	1	71	21	0	0	0	0	0	0	2
	bit	0	0	3	3	1	29	55	7	0	0	0	0	0	2
	bought	0	0	0	0	0	0	0	0	59	3	7	22	8	1

Table 5.4: Confusion matrix showing the vowels (bet, bit, bought) identified by SSBE listeners (percent correct), for the three training groups at the post-test

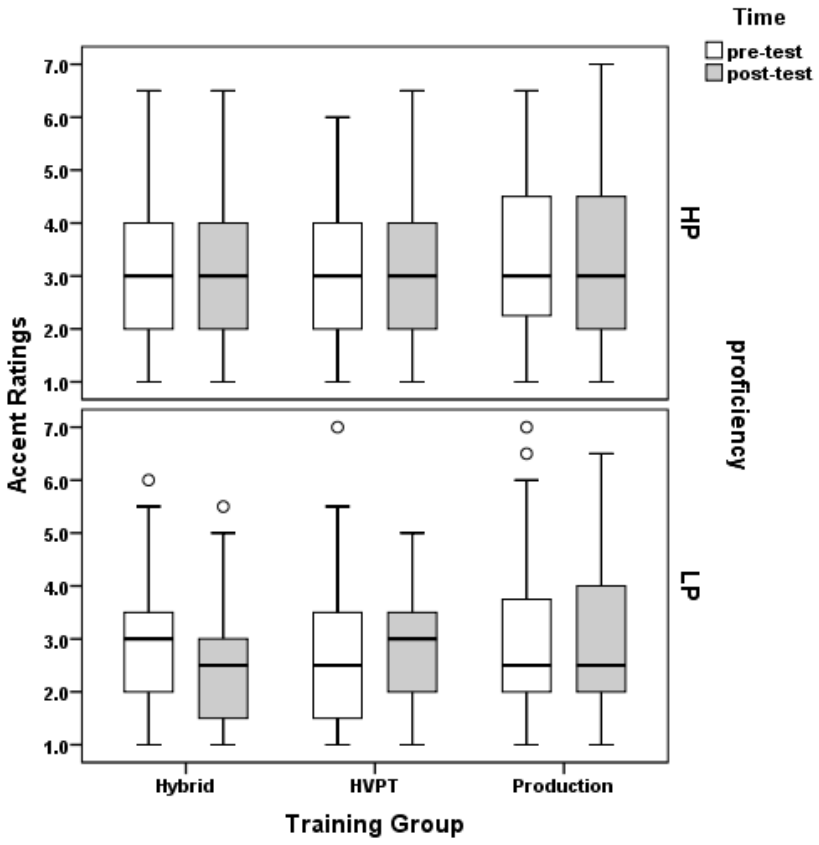


Figure 5.16: Boxplots for goodness ratings for the IEEE sentence "*Glue the Sheet to Dark the blue Background*" at the pre- (white boxes) and post-test (grey boxes), split by training group and proficiency (HP = top panel, LP = bottom panel). Sentences were rated on a Likert scale from 1-7, where 1= strong-accent, 7 = close to natives' production.

Goodness Ratings. As displayed in Fig.5.16, overall, L2 learners were rated as having strongly accented-speech (i.e., they received a low rating score at both the pre- and post-test). Additionally, there appeared to be no effect of training on SSBE listeners' ratings. In order to investigate whether the ratings were reliable, a reliability test was run using the Intra-class Correlation Coefficient (ICC) on the 10 raters' scores for the snippets from the pre- and post-test. To test the level of rater-agreement, a two-way mixed model was chosen with "Absolute Agreement" type and raters as fixed components (i.e., whether the raters used the scale in the same or similar way). The results demonstrated a strong consistency in the ratings, Cronbach's Alpha α=.837 (a perfect Cronbach's Alpha=1).

After confirming that the ratings were reliable, a linear mixed effects model was built for the rating scores. The best-fitting model included time (pre-post), and training group (PT, HVPT, and HTP) coded as fixed factors, and participant (rater) and speaker coded as random factors with random slope. Proficiency was excluded indicating that this was not a significant factor. The results from the model showed that there was no significant effect of time which suggests that there was no significant change in accent ratings before and after training. There was also no significant effect of training group, indicating that there was no significant difference in accent ratings across training groups. However, there was a significant two-way interaction between time and training group, χ^2 (2) = 7.336, $p < .05$. The planned contrasts indicated a significant difference from pre- to post-test for the HVPT group, b= -0.305, SE= 0.1203, pMCMC<.05, also a significant effect from pre- to post-test for the PT group, b= -0.252, SE= 0.118, pMCMC<.05.

5.2.5 Links between production and perception

A series of Pearson correlations investigated whether or not there was a link between vowel identification (i.e., Arabic learners identifying SSBE vowels) and vowel intelligibility (i.e., SSBE listeners' identification of Arabic learners' English vowels).

Figure 5.17 displays the relationship between L2 learners' pre-test vowel identification scores, and their pre-test vowel intelligibility at the pre-test. A Pearson's correlation indicated a significant correlation between vowel identification and vowel intelligibility at the pre-test, [r=.675, p<.001, R2= 0.455]; participants who performed better on the vowel identification task also tended to be more intelligible.

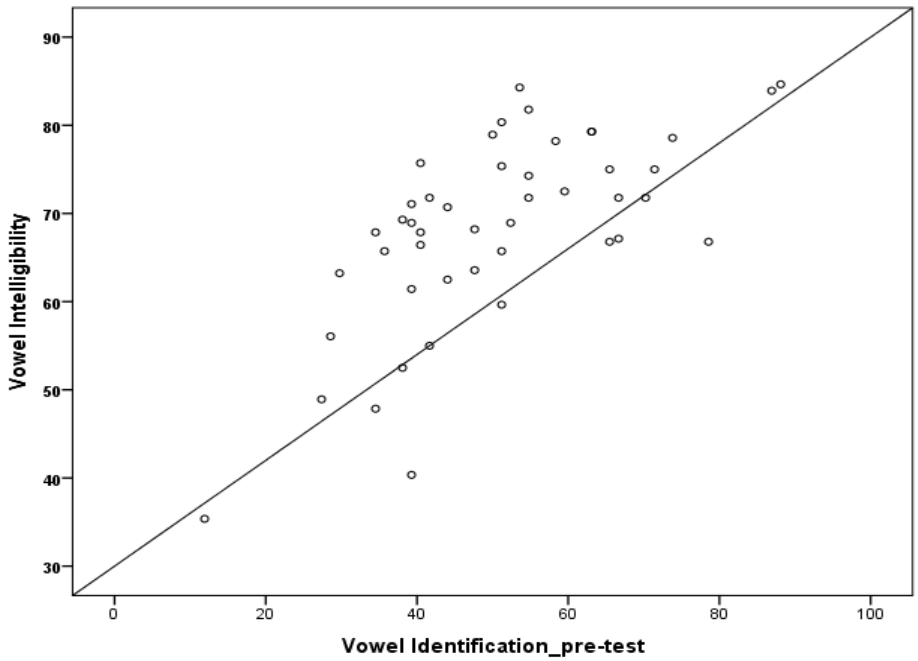


Figure 5.17: Scatterplot showing the correlation between vowel intelligibility and vowel identification at the pre- test

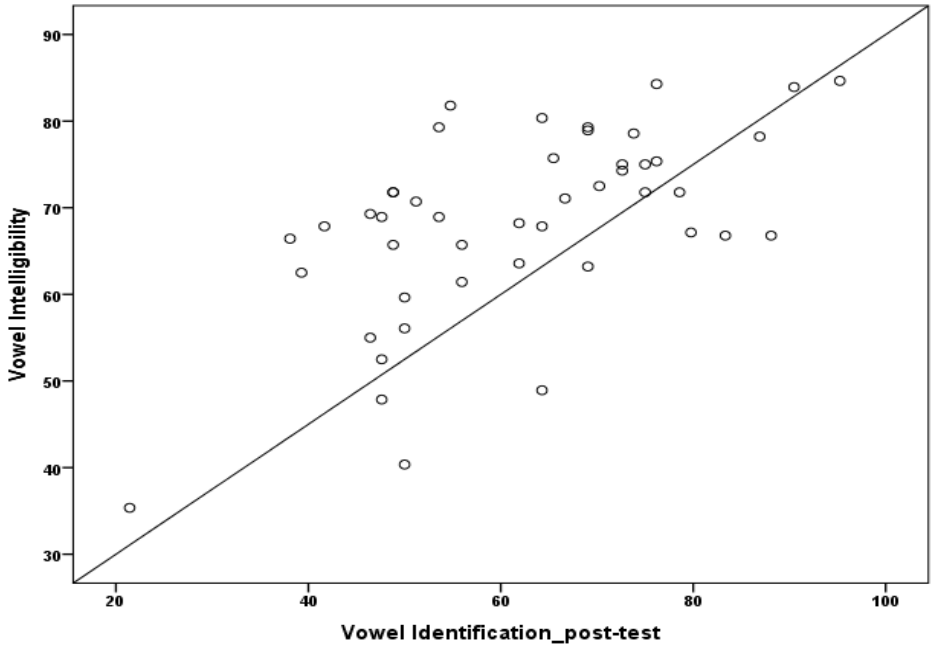


Figure 5.18: Scatterplot showing the correlation between vowel intelligibility and vowel identification at the post-test in percentage.

Fig. 5.18 displays the relationship between L2 learners' post-test vowel identification scores and their post-test vowel intelligibility at the post-test. As for the pre-test, a Pearson's correlation showed a significant correlation between vowel identification and vowel intelligibility at the post-test, [r=.599, p<.001, R2= .358] such that participants who performed better on the vowel identification task also tended to be more intelligible.

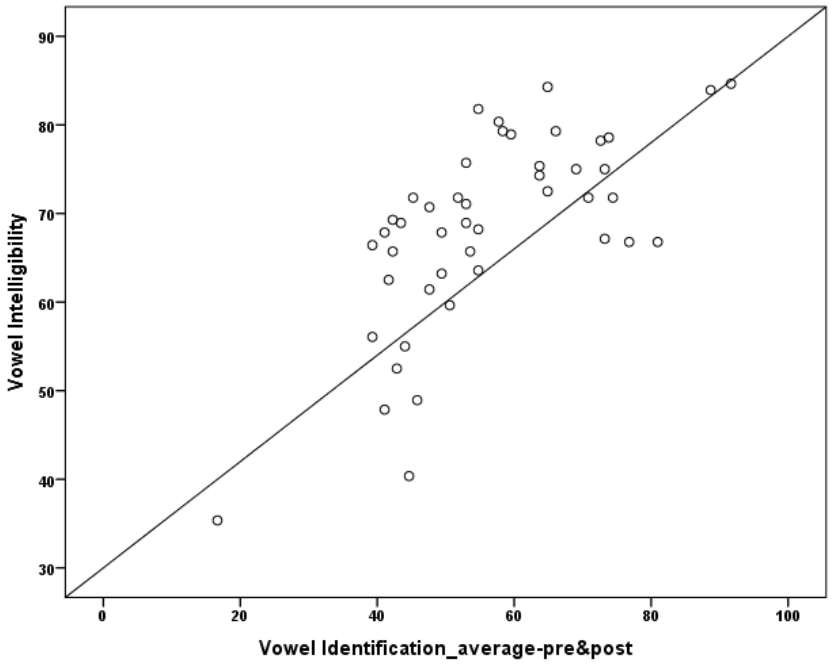


Figure 5.19: Scatterplot showing the correlation between vowel intelligibility (the average between the intelligibility at the pre- and the post-test) and vowel identification scores (the average of the vowel identification tasks at the pre- and the post-test).

Fig. 5.19 displays the relationship between the average vowel identification scores (averaged over the pre- and post-test) and average vowel intelligibility scores. As expected, a Pearson's correlation indicated a significant correlation between vowel identification and vowel intelligibility, [r=0.679, p<.001, R2=.461], indicating that performance on a perception task was an indicator of production accuracy for all

groups. That is, overall, L2 learners who demonstrated accurate vowel identification, produced accurate vowels, and therefore, were highly intelligible to the SSBE listeners, whilst those who had lower vowel identification accuracy were less intelligible.

The correlation between the average vowel intelligibility and average vowel identification scores was also similar across training groups. Individual Pearson correlations for each training group showed a significant correlation between vowel identification and vowel intelligibility with similar R2 values; PT [r=.652, p<.05 =.006, R2= 0.42]. HT [r=.759, p<.05=.001, R2 =0.57] and HVPT [r=.606, p<.05 =.017, R2= 0.36].

5.3 Discussion

The present study examined the effect of three different training programs, PT (production based), HVPT (perception-based), and HTP (production and perception) on vowel production and perception by Arabic learners of English. In particular, the study aimed to investigate whether phonetic training for second language learning is domain specific (i.e., whether PT leads to improvements in both production and perception, or just perception). The results demonstrated that different types of training affected performance in production and perception tasks differently. After training, learners who had completed perception-based training programs (HVPT and HTP) improved more in their vowel identification than those who completed PT. However, those who received some training in speech production (PT and HTP) improved more in production that those who received only perception training (HVPT). Additionally, the results demonstrated that initial proficiency in the L2 affected learning in some tasks, in particular, speech in noise.

Overall, these findings indicate that training is largely domain-specific; that is, production trains production and perception trains perception. Previous research has shown that HVPT is particularly effective in improving perception (e.g., Logan et al. 1991; Iverson & Evans, 2009) and learners in the HVPT and HTP training conditions also improved significantly more in their vowel identification than did those who

received PT. This may be because HVPT enabled learners to become better and more efficient at mapping their native categories onto the L2 sounds they heard but without necessarily making changes to their underlying representations (see also Iverson & Evans, 2009). Similarly, PT may have enabled learners to develop more native-like production for particular vowels that they were able to map onto their existing underlying representations. That is, they may have learned new motor commands that they were able to map to their existing representations but they may not have altered the underlying representations themselves.

These findings are in contrast to previous research which has found that improvements in perception as a result of HVPT training generalized to production (e.g., Bradlow et al., 1997). Learners in the HVPT condition did not improve in production, only those who received production-based training (PT & HTP) produced more native-like vowels after training. Learners in both these conditions were also able to generalize their production to a different set of words (i.e., not included in the training), and adjusted their production of some confusable vowels, in particular /ɪ/ and /e/. Before training, learners produced /e/ as a more closed vowel (i.e., more similar to /ɪ/), and /ɪ/ as a more open vowel (i.e., more similar to /e/). One explanation for this could be that initially, their native vowel space made it difficult for them to distinguish these categories. Since the Arabic vowel inventory (i.e., /i, iː a, aː, u, uː/) does not include either of the English vowels /ɪ/-/e/, Arabic learners likely find it difficult to produce or perceive these two "novel" vowels as two different categories. Consequently, they might establish a single new category that is close to their closest matching L1 category (i.e., /i/), and which includes the two English vowels /ɪ/-/e/. That is, though they may have been aware that there was a difference between /ɪ/-/e/, (e.g., as a result of orthographic cues), the initial formation of this new category was likely not robust enough to enable Arabic learners to distinguish the two vowels reliably (SLM; Flege, 1995, 1999, 2002).

After training though, mostly all participants produced these vowels more like native speakers, such that /ɪ/ was produced with a lower F1 than /e/. Such improvement provides some evidence that explicit instructions and feedback with visual

representations of the lips, tongue and jaw are effective for improving vowel production.

However, although PT led to improvements in the production of the vowel contrast *bit-bet* contrast, other vowel contrasts did not improve significantly after training. In particular, learners did not change their production of *boot*, or improve in their production of the *bot-bought* contrast. In SSBE and other accents of British English, /u/ is typically fronted, such that it 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 it to their native category [u] (PAM: Best, 1995; Best &Tyler, 2007). This type of assimilation is known as a single category assimilation in PAM (Best, 1995; Best &Tyler, 2007), and in these cases where L1 and L2 categories are assimilated equally well or poorly to a single L1 category, the discrimination is predicted to be very poor. This in turn, might have prevented learning in production. Additionally, participants may not have been motivated to change their pronunciation of this vowel. Although the 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 be understood, and that native English listeners would not be likely to assimilate the high rounded back vowel variant to their high central rounded variant.

In contrast, L2 learners may not have improved in their production of the *bot-bought* contrast because these two vowels do not exist in their L1 vowel inventory. Consequently, their native vowel space may have biased them to hear differences in a particular way so they could not easily relate differences between sounds as a result of the new articulatory patterns they have learned to the differences that they heard. This may indicate that production training necessarily involves perception (i.e., learning to relate new motor patterns to perceived 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.

Additionally, one might interpret the lack of improvement in the back vowels, as a result of the fact that there was just too much information over the 5 sessions for the L2 learners to process. Therefore, learners may have focused their attention on vowels that they thought were more difficult for them (i.e., /ɪ/, /e/). It is possible that with more sessions, they may have been able to change more aspects of their production.

One limitation of the current study is that learners were not recorded within the sessions. Such recordings would have enabled vowel production to be tracked during each session and for the rate of improvement to be measured. For instance, if learners had improved gradually from the first to the fifth session, that would indicate they still had scope for improvement after the fifth session, and thus, five sessions were not enough. While if learners improved from the first to the third session, but then the amount of improvement plateaued, it would indicate that learners have reached a ceiling and would not have greatly benefitted from further sessions using the same training technique.

That being said, the data from the confusion matrices for the vowel intelligibility showed a change in one back vowel, *bought*, which improved 26% in PT learners and 19% in the HTP learners, but only 9% in HVPT learners. SSBE listeners also identified *boot* quite accurately even if it is far acoustically from English [ʉ]. This could be because native listeners had learnt something about Arabic vowels during the identification task that made *boot* intelligible, even if it is not produced similarly to English [ʉ]. That is, they had adapted to non-native version of *boot*.

It was surprising that production did not lead to improvements in perception as the design of the articulatory training meant that learners listened to examples of keywords, low frequency real words, and isolated vowels, as well as their own and the instructor's examples. This means that they were perceiving speech as well as

receiving articulatory instruction, yet, their vowel perception did not improve after training. This is possibly because any production-based learning within the five sessions did not yield robust enough L2 category learning for all vowels for that knowledge to be transferred to the other speech domain (i.e., perception). That is, if we assume that perception and production share the same underlying categories, learners may have acquired new motor commands and may have linked these to their underlying representations, but may not have made any changes to the underlying representations themselves, with the result that the learning does not pass across to perception. Perhaps, the transfer from production learning to perception might take longer to occur. That is, having completed the five sessions of production training, learners might be able to use the skills they have acquired to maintain and possibly build on improvements in production, and these improvements may subsequently pass to perception. This would necessarily require further exposure to and use of English. This question is investigated in Chapter 6 in which learners who had remained in the UK were re-tested 6 months after training.

Likewise, perception training did not lead to improvements in production. Although this seems more intuitive (HVPT did not involve any production training) there is some anecdotal evidence that some learners may actively attempt to reproduce what they hear even when doing perception task. That is, while listening to word stimuli, they may have practised producing the words they heard, or repeated minimal pairs when they got corrective feedback from the program in order to help them distinguish between certain vowel contrasts. Additionally, based on the fact that some studies found that when perceiving speech, the motor areas involved in speech production are activated (e.g., Wilson et al., 2004), one might assume that L2 learners' production would improve after the HVPT. This assumption might be reasonable since speech perception and production have been shown to have a strong link.

Such a link between perception and production was confirmed by a positive correlation between L2 perception (i.e., how accurately L2 learners identify English vowels) and L2 production (i.e., how accurately SSBE listeners identify vowels produced by L2 learners), replicating previous studies (e.g., Flege, 1993; Bradlow,

1997). This correlation may suggest that some aspects of perceptual knowledge may be related to L2 production. However this correlation was only found between L2 learners' vowel identification score and vowel intelligibility (i.e., native listeners' perception of the L2 learners' vowels).

So why no improvement in production after HVPT? One possibility is that, the effect of training might differ according to the tests that measure the learning effect of one speech domain or the other. That is, the vowel identification task was the means of investigating improvement in perceptual abilities, and thus participants who received HVPT performed better. This may be because the HVPT is based on identification with corrective feedback, and the vowel identification task is also an identification task but without correction. Consequently, there is a possibility that the HVPT helps learners to be better at identifying the target words or phonemes but only in this particular type of task. However, their vowel production did not improve, because production was not emphasized in the HVPT. A similar suggestion could be made about the production training. Training L2 participants on production perhaps re-directed their attention to produce a certain vowel in a particular way, and this might led to only surface changes in production, rather than changes which led to changes in underlying category representations.

Another explanation is that learners need explicit training that directs or re-directs their attention to the trained method or materials, so that they can attend to certain acoustic cues. Such explicit training may help shifting their attention from the trained materials to generalize the obtained knowledge during training to untrained phonetic cues (cf. Francis et al., 2000). That is, in the current study, learners who were trained using perceptual training had their attention directed to be better at identifying phonemes but not necessarily to become better at producing them. The same can be said about training production; production training directed learners to produce native-like phonemes, but not necessarily to perceive them more accurately.

The results from the HTP programme provide some support for the importance of explicit feedback in both perception and production. This programme included only

one production training session (besides 4 sessions of the HVPT), but the results showed that L2 learners in this training condition improved in both production and perception. Although these learners did not improve in production as much as those as in PT, this suggests that only a small amount of production training may be needed to effect some change to production. Consequently, training L2 learners in both perception and production may be the best approach. Further studies might thus consider including an equal number of perception and production training sessions to investigate whether both speech domains would improve even further.

What is being learned in production and perception training? Interestingly, all participants improved in their performance on a speech recognition in noise task, regardless of training type. This suggests that the finding that learning is domain-specific may in part be task driven. In HVPT programs like that used in the HVPT and HTP training conditions, learners identify different sets of minimal pairs, the same skill tested in the vowel identification task. Given that there was very little evidence for changes in low-level speech perception (i.e., few changes in performance on a Category Discrimination task), this indicates that learners were not making changes to underlying representations as a result of training. Instead, it is possible that HVPT enabled learners to become better and more efficient at mapping their native categories onto the L2 sounds they heard (see also Iverson & Evans, 2009).

In contrast, improvement in a more real-world task of speech perception, sentence recognition in noise, appeared to rely on initial proficiency with the L2 rather than training type. Although all learners improved in their performance in speech in noise, HP learners improved more than LP 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 then is that a certain level of grammatical and lexical knowledge is necessary to apply learning on isolated sounds and words to real-world contexts.

In brief, the results from this study confirm that phonetic training is largely domain specific and additionally indicate that adjustments to phonetic processing

might be lexically driven (i.e., certain level of knowledge is necessary to apply learning on isolated sounds and words to real-world contexts, as HP participants in speech recognition in noise task). Perceptual training led predominantly to improvements in speech perception, whilst production training, even only a small amount, led to changes in production. However, performance on a speech in noise task was affected predominantly by proficiency rather than by training type. 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 might differ.

Chapter 6 Investigating the long-term retention of learning in perception and production

6.1 Introduction

In the previous chapter it was argued that phonetic training is domain specific, and that though perception and production are likely linked in some way, this link may not be as direct as might have been previously assumed (e.g., Bradlow, 1997; Wang et al, 2003) at least for the effect of the training of one speech domain on the other. This study investigates whether training has long-term effects on perception and production, and whether different training types (HVPT, PT, and HTP) yield differences in long-term learning retention.

Previous research into phonetic training (e.g., Bradlow, 1997; Iverson & Evans, 2007) has found that HVPT is successful in improving speech perception or/and speech production (see p. 82 for review), and that learners maintain their improvement after completing training (e.g., Lively et al, 1994; Bradlow et.al, 1999; Iverson and Evans, 2009). For example, Iverson and Evans (2009) tested Spanish learners immediately after they had completed 5 sessions of HVPT, and then again 2-6 months later. Likewise, German speakers were tested immediately after completing 5 sessions of HVPT and then one year later. Both Spanish and German learners showed improvement in their vowel identification performance immediately after training, and learners retained these improvements up to one year post training.

Likewise, there is evidence for retention of learning from CALL-based training programmes. For example, Wang and Munro (2004) trained Mandarin and Cantonese speakers on the perception of three English vowel contrasts (/i/-/ ɪ /, /u/-/ʊ/, and /ɛ/-/æ/) over a period of two months. Learners completed identification tasks for synthetic and natural /h/-V-/d/ tokens. They were tested immediately after completing training and then again 3 months later. The results demonstrated that learners improved their vowel identification, generalised this learning to new tokens and new speakers after completing training, and that identification performance was the same 3 months later, indicating that learning was not lost.

However, all of these studies focused on the retention of perceptual training, and only tested whether or not improvements persisted for performance on speech perception tasks. To my knowledge, there is no research investigating the retention of learning for production training or retention of improvements in speech perception and production as a result of production-based training.

The current study investigated retention of learning in participants who had completed the training Study 1 (see Chapter 5). This included participants from the two production-based training programmes, PT, and HT, as well as those who had completed perception-based training (HVPT). All participants were contacted 6 months after completing the initial training, and 22 took part in the retention study, Study 2. These participants completed a sub-set of pre- and post-tests used in the initial training study; vowel identification, vowel production, and speech recognition in noise. For practical reasons, the category discrimination task was not included as the initial training study had demonstrated no robust effect of training on performance in this task.

6.2 Method

6.2.1 Participants

Twenty-two participants from Study 1 (Chapter 5) participated in this study. Their ages ranged from 20-38 years (median 29 years old). Participants were resident in the UK at the time of testing and had 3- 69 months of experience living in an English speaking country (median 3 years). Almost all subjects reported more daily interactions with speakers from their home country and non-native English speakers of other language backgrounds than they did with native speakers of English. The number of the subjects in each group was dependent on the ability to re-contact them. In total, 8 participants in the PT group, 8 from the HT, and 6 from the HVPT group took part. All subjects reported no speech/hearing problems.

6.2.2 Apparatus

This was the same as Study 1 (Chapter 5). In brief, tests were conducted in a quiet room with stimuli played over headphones (Sennheiser 555) at a user-controlled comfortable level. Stimuli were played via a Dell Inspiron N5040 laptop with built-in sound card. The same PC laptop was used to collect responses via an experimental interface. 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.

6.2.3 Stimuli

Participants completed a subset of the pre- and post-tests used in Chapter 5; vowel identification, speech recognition in noise and vowel production. The stimuli were the same as those used in Study 1. In brief, they identified /b/-V-/t/ words in quiet and IEEE sentences in noise, and were recorded producing the /b/-V-/t/ words they had identified in the vowel identification task (see 98 for details)

6.2.4 Procedure

All participants were tested in quiet rooms. They first completed the perceptual tasks (vowel identification and speech recognition in noise) before recording the bVt words. The procedure for each task was the same as in Study 1 (see p.106 for more details)

6.3 Results

6.3.1 Vowel Identification

As shown in Figure 6.1, some learners appeared to have retained any improvements in vowel identification, although there appeared to be some effects of training type and proficiency (see Fig. 6.2). HP learners who had completed PT, HVPT and HT training performed at the same level at the retention test as they did at the initial post-test. LP learners on the other hand performed similarly in the retention test to how they did in the post-test in both the HTP and HVPT training conditions, but

those in the PT condition showed further improvement from the post-test (mean score = 0.47) to the retention test (mean score = 0.51).

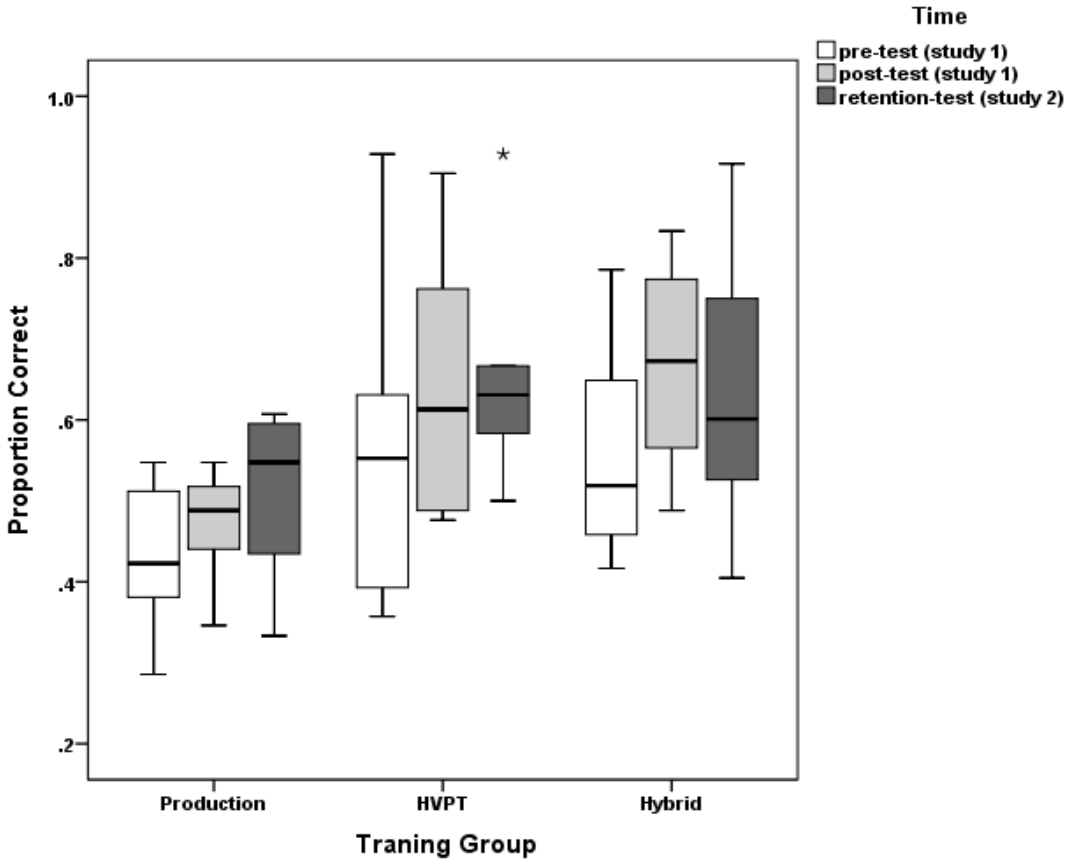


Figure 6.1: Boxplots of the proportion correct in the retention test compared to pre- and post-test scores from Study 1 (Chapter 5), for the three training groups (PT, HVPT, and HTP). The pre- and post-test scores include data from only those participants who completed both Study 1 (Training) and Study 2 (Retention).

To verify any changes in vowel identification performance across the three testing sessions (pre-, post- and the retention test), a logistic mixed effects model was built for the identification analysis based on binomial responses (correct/ incorrect). The best-fitting model was chosen with a top-down approach (i.e., excluding ineffectual random and fixed factors from a model with all potential factors). The best-fitting model included time (pre, post, retention), training group (PT, HVPT, HTP), proficiency (HP, LP), and the interaction between group and proficiency coded as fixed factors. Participant and stimulus were included coded as random factors with a

random slope for each of pre-, post, and the retention tests. The model excluded the interactions between time and training group, time and proficiency, and time, training group and proficiency, which means that these interactions are not significant.

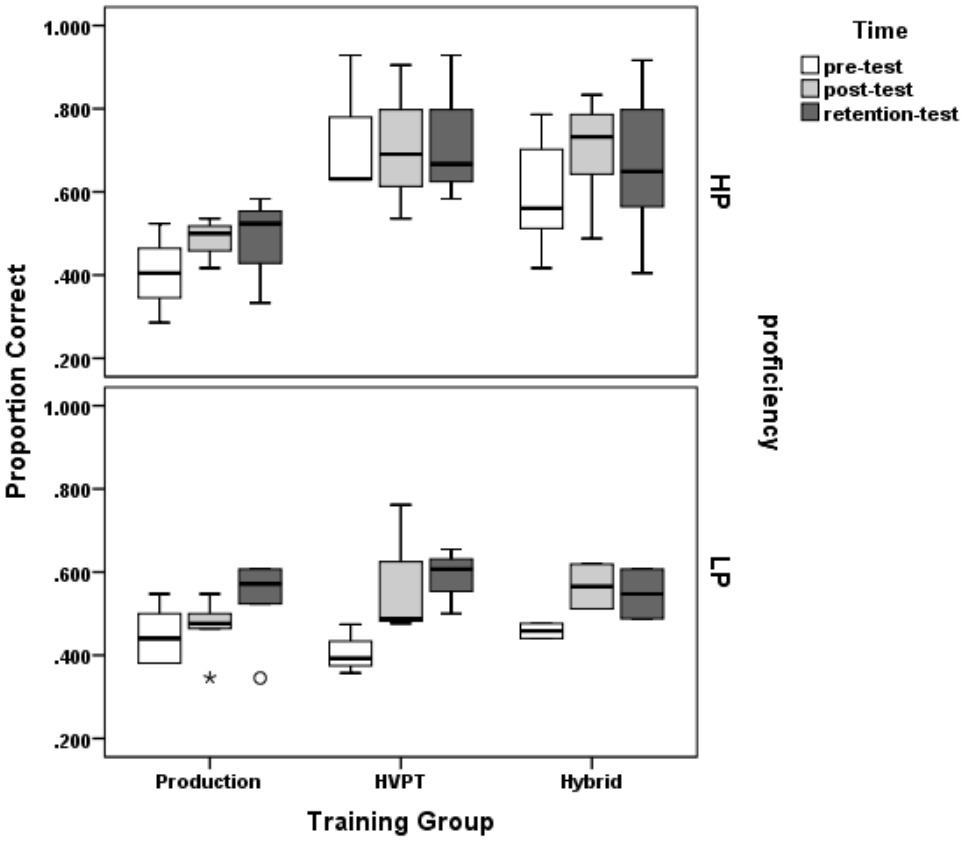


Figure 6.2: Boxplots of the proportion correct of the vowel identification task in the retention test compared to pre and post-tests from Study 1 (Chapter 5), split by proficiency level (HP, LP) for each training group (PT, HVPT & HT).

The results from the model demonstrated a significant effect of time (pre, post, retention), χ^2 (2) = 17.755, p < .0001, confirming that learners improved in their vowel identification. As expected, the planned contrasts indicated a significant improvement from pre- to post-test, b= -0.2803, SE= 0.068, z= -4.068, p<.0001. The planned contrasts also indicated a significant improvement from post-test to the retention test, b= -0.1659, SE=0.0794, z=-2.089, p<.05. There was a significant effect of training group, χ^2 (2) = 7.2991, p < .05. The planned contrasts indicated no significant difference between the PT and HT groups, or between HTP and HVPT groups.

However the difference between the HVPT and PT groups just reached significance, b= 0.2955, SE=0.1530, z=1.931, p=.05, indicating that there was a marginal difference in performance between the groups in the vowel identification in the PT group compared to that of the HVPT group; the PT group performed more poorly overall than did the HVPT group.

There was a significant effect of proficiency, χ^2 (1) = 8.0701, p < .05, confirming that the proficiency level of L2 learners affected performance. As expected, the planned contrasts indicated that overall the HP group scored higher than the LP group, b= 0.3285, SE= 0.1058, z= 3.103, p<.001 (see Figure 6.2). There was also a significant interaction between training group and proficiency, χ^2 (2) = 7.494, p < .05, suggesting that learners with different proficiency levels performed differently according to training group. The planned contrasts indicated that the HP group who received HVPT performed better overall than HP learners in the PT training condition, b= 0.355, SE= 0.1530, z= 2.326, p<.05. However, there were no differences in performance between LP participants in the HVPT and PT training conditions.

6.3.2 Speech recognition in noise (IEEE-sentences)

Figure 6.3 displays the box plots for the speech reception threshold (SRT) for the participants at each different testing time (pre, post and retention tests). As shown in Fig 6.3, there appeared to be a change in the SRT level from pre- to post-test, and from post-test to the retention test; the majority of participants appeared to have improved in their speech recognition in noise at both the post-test and retention test.

To verify any significant changes, a linear mixed-effects model was built using top-down procedure as described above. The best fitting model included time (pre, post, and retention), training group (PT, HVPT, HTP), and proficiency (HP, LP) as fixed factors, and participant as a random factor with a random slope. The results from the model indicated a significant effect of time, χ2 (2) = 126.112, p<.001, confirming a change in the SRT from pre- to post and/or from post- to the retention test.

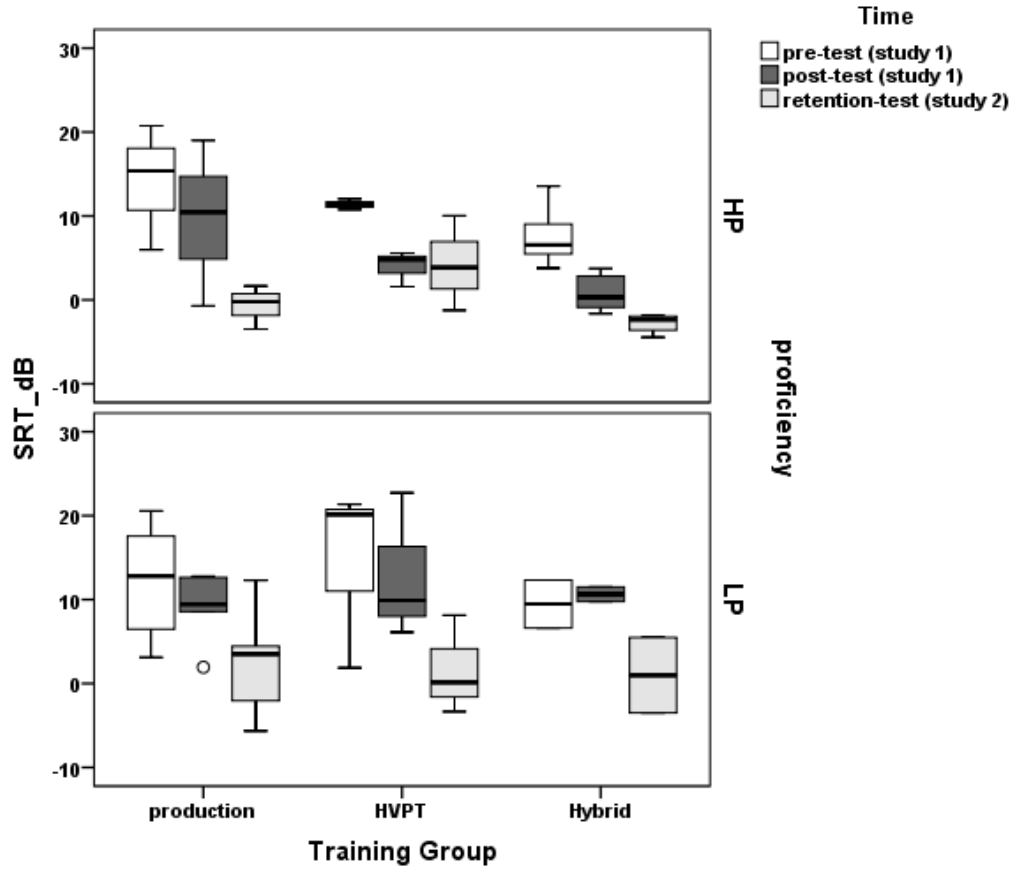


Figure 6.3: Boxplots of the speech reception threshold (SRT) for L2 learners at pre, post (Study 1, Chapter 5), and the retention tests, split by proficiency level (HP &LP) for each training group (PT, HVPT, HT). The pre- and post-test scores include data from only those participants who completed both Study 1 (Training) and Study 2 (Retention).

As expected, based on Study 1, the planned contrasts between different testing times showed a significant change in SRT from pre- to post-test; participants performed better at the post-test than at the pre-test, b=4.7245.639, SE= 0.579, pMCMC<.001. The planned contrasts also confirmed that performance improved from the post-test to the retention test, b=-5.784, SE= 0.5791, pMCMC<.001.

There was no significant effect of proficiency, but there was a significant interaction between time and proficiency, $\chi 2$ (2) = 6.589, p<.05, suggesting that participants with different proficiency levels improved at different rates from pre- to

post-test, and from the post-test to the retention test. As displayed in Fig 6.3, LP learners appeared to improve more than HP learners from the post-test to the retention test. The planned contrasts indicated no significant difference at the retention test between LP and HP participants. However, there was no significant difference between the two proficiency groups from pre- to post-test. This is in contrast to the results in Study 1 which showed that HP but not the LP learners improved in this task from pre- to post-test. This may be because in this study, we tested a subset of the participants in Study 1. There was a significant three-way interaction between time, training group, and proficiency, $\chi 2$ (2) = 11.589, $p<.05$. The HP learners in both the HVPT and the HTP groups improved from pre-to post-test, while the PT did not improve from pre- to post test, but they did improve from post to the retention-test.

The LP learners appeared to improve from pre-to post-test in both the PT and HVPT, while the HT learners did not. All LP learners in all training groups appeared to improve at the retention-test (see Figure 6.3).

6.3.3 Vowel production (/b/-V-/t/ words)

As in Study 1, in order to avoid multiple comparisons, the monophthongs were divided into three groups; (beat, bit, bet, bert), (bat, but, bart), and (boot, bought, bot). The analysis of F1& F2 for each vowel group is presented first, followed by an analysis of duration, again, for each vowel group. To enable comparison of male and female talkers, formant frequency measurements were normalized using Lobanov's method (Lobanov, 1971).

6.3.3.1 Spectral analysis

Group 1: Beat, Bit, Bet, Bert. Figure 6.4 shows F1 and F2 measurements at the pre- post- and retention tests. They show some change in F1 and F2 values across the 3 testing sessions. Specifically, Study 1 showed that learners in the PT group changed their F1 values for the *bit-bet* contrast from the pre- to post-test, such that at the post-test *bit* was produced with a higher F1 and *bet* with a lower F1 to better match native speakers. This learning appears to have been retained from the post- to retention test.

In order to test these observations, a linear mixed effects model was built for F1 and F2 separately. The best fitting-model for F1 for this vowel group included time (pre, post, and retention), training group (PT, HVPT, HTP), and proficiency (HP, LP) as fixed factors, and participant and stimulus as random factors with a random slope for time. The model excluded the interaction between time and proficiency, and the interaction between time, group, and proficiency, which indicates that these interactions are not significant for the analysis.

The results from the model indicated no significant effect of time, suggesting no significant change from pre- to post test, and no change from post-test to the retention test. There was no significant effect of training group (i.e., there was no overall difference in the F1 values used by participants in the different training groups). However, there was a significant two-way interaction between time and group, χ^2 (4) = 16.297, $p<.05$, suggesting that some training groups changed their F1 values across time. As expected, the planned contrasts showed a significant change in F1 from pre- to post-test in the PT group compared to the HTP group, b= -0.1623, SE= 0.0496, pMCMC<.05 (see Figure 6.4).

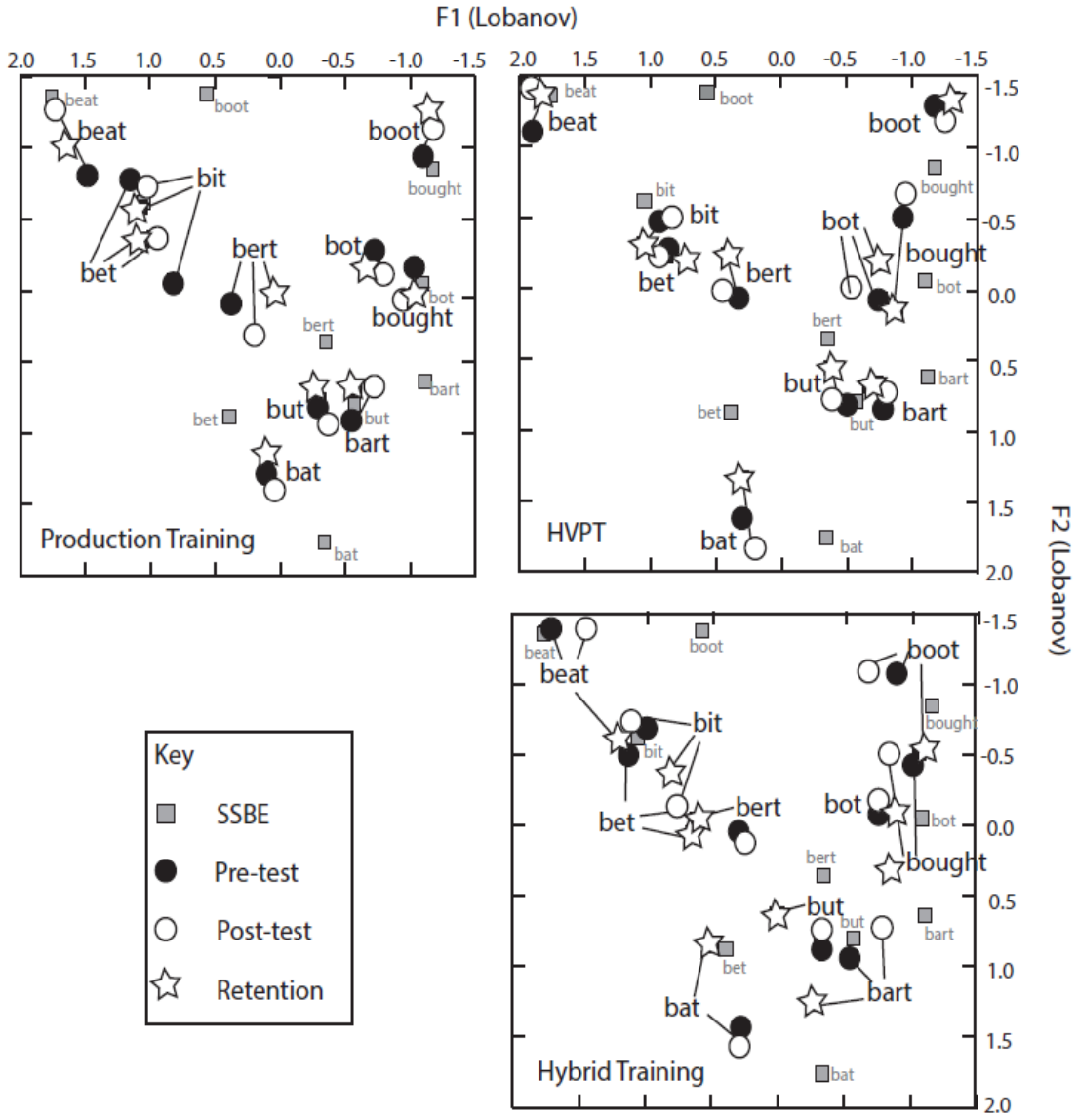


Figure 6.4: Vowel plot showing the vowel space of L2 learners at the pre-, post- and the retention test, compared to those of the SSBE. F1 and F2 values were normalized using Lobanov method.

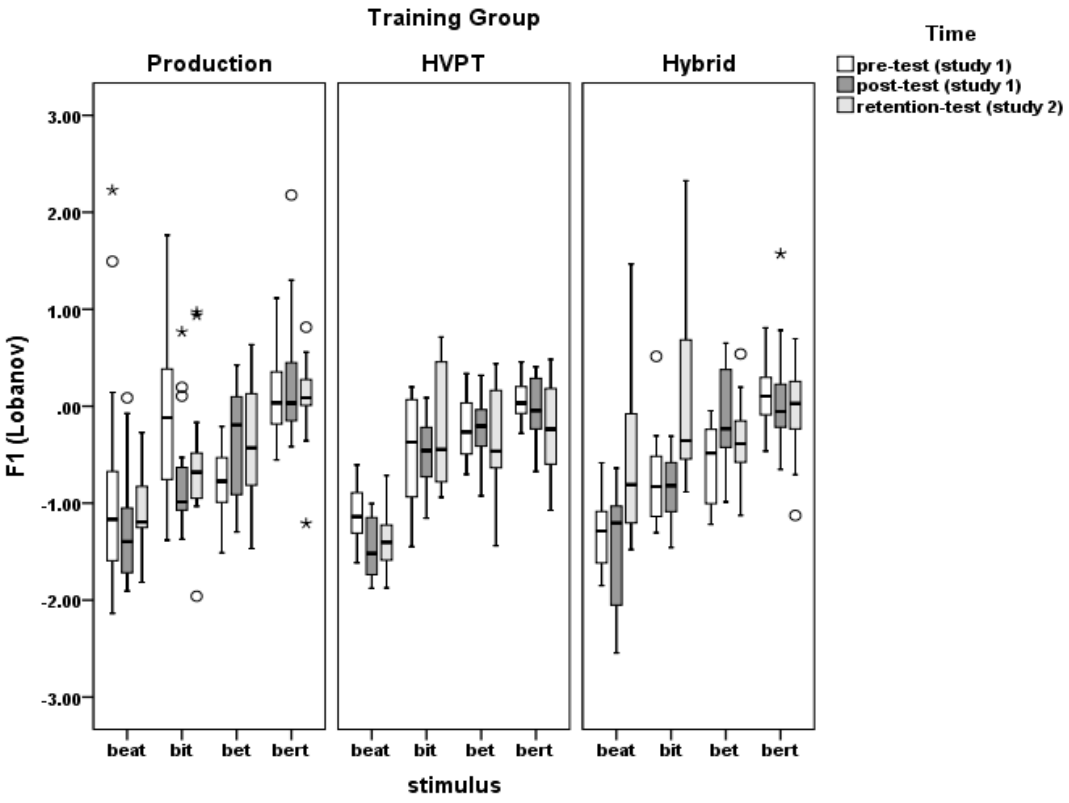


Figure 6.5: Boxplots of F1 for Vowel group 1(beat, bit, bet, bert) produced by L2 learners in the 3 training conditions (PT, HVPT, HTP). Formant values were normalised using Lobanov's method

Although there was a wide standard deviation (Std. =.705), participants in the PT group tended to produce *bit* with lower F1 values, and *bet* with higher F1 values than they did at the pre-test, such that their production of the *bit-bet* contrast better matched that of native speakers (see Figure 6.4).

There was also a significant change in F1 values from post-test to the retention test; participants in the HT group changed their F1 value more than those in the PT group, b= 0.1834, SE= 0.04997, pMCMC<.001. Learners in the HT group produced *bit* with a higher F1 at the retention test whereas learners in the PT group produced *bit* with lower F1 values. There was no significant difference in F1 values between the

learners in the HVPT condition from pre- to post test, or from post-test to the retention test, pMCMC>.05.

For F2 values, the best fitting-model included time (pre, post, and retention), and training group (production, HVPT, hybrid) coded as fixed factors, and participant and stimulus coded as random factors with random slope. The results from the model indicated no significant effects of the factors, indicating no significant change in F2 values (see Figure 6.4).

Group 2: Bat, But, Bart. As shown in the Figure 6.4, there appeared to be little change in F1 or F2. In order to test these observations, a linear mixed effects model was built for F1 and F2 values separately.

For F1, the best fitting model included time (pre, post and the retention test), and proficiency (HP, LP) coded as fixed factors, and participant and stimulus coded as random factors with random slopes. The results from the model showed no significant effect of the factors, confirming that there was no significant change in F1 values for this vowel group.

For F2 values, the best fitting-model included time (pre, post, retention), training group (PT, HVPT, HPT) and proficiency (HP, LP) coded as fixed factors, and participant and stimulus coded as random factors with random slopes. The results from the model indicated a significant effect of time, χ^2 (2) = 23.6541, $p < .001$, suggesting a change in F2 values from pre- to post and/or from post-test to the retention test. The planned contrasts showed a significant change in F2 values from post-test to the retention test, b= 0.4673, SE= 0.0859, pMCMC<.001. however, there was no significant change in F2 values from pre- to post-test p>.05, contradicting the findings from Study 2, where participants changed their F2 values from pre- to post-test. The effect of training group was also significant χ^2 (2) = 7.6989, p<.05, indicating differences in F2 values between training groups. However, the planned contrasts showed no significant differences between the different groups, which suggests that

although there was more variability in the HPT than in the PT and HVPT groups, there was no reliable difference in production across training groups.

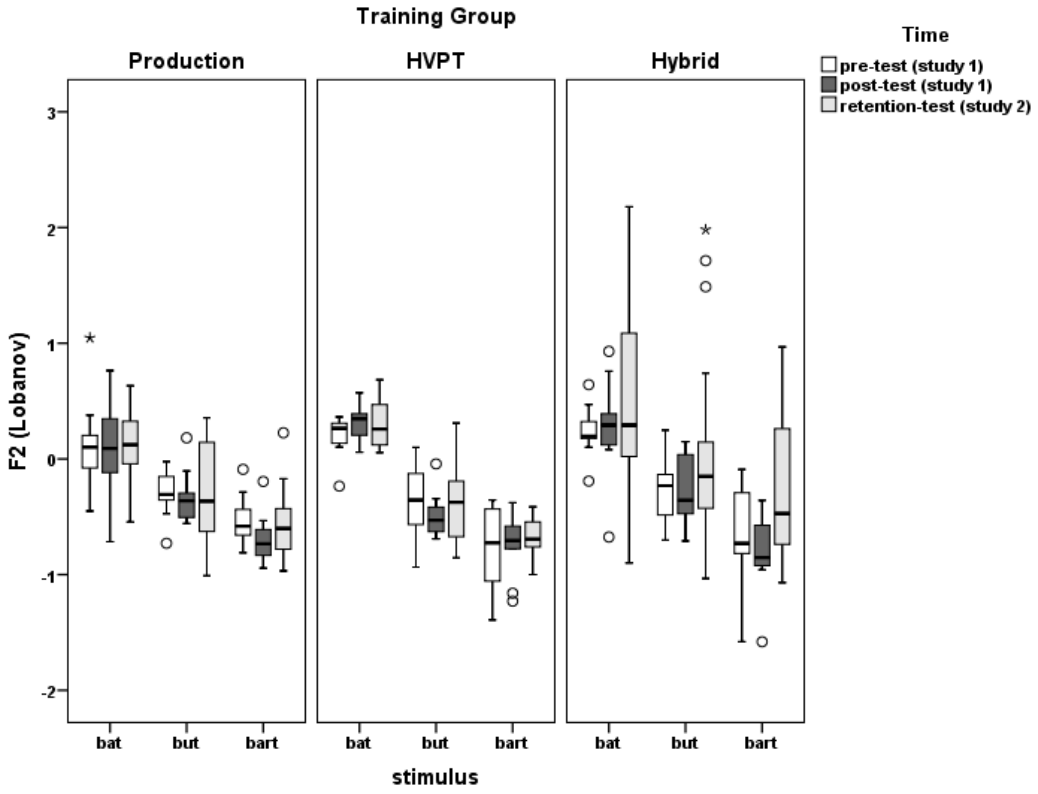


Figure 6.6: Boxplots of F2 for Vowel Group 2 (bat, but, bet, bart) produced by L2 learners in the 3 training conditions (PT, HVPT, HT). Formant values were normalised using Lobanov's method.

However, there was a significant two-way interaction between time and group, χ^2 (4) = 15.6950, p< .05. The planned contrasts showed a significant change in F2 values from the post-test to the retention test for the HVPT group b= -0.3584, SE= 0.13586, pMCMC<.05; in particular, this group changed their production of *but* at the retention test such that it was similar to the pre-test, (Figure 6.6). The planned contrasts also showed a significant change in F2 values from post-test to the retention test for learners in the PT group, b=-0.5449, SE= 0.135, pMCMC<.001; in particular, learners changed their production of *bart* at the retention test, such that it was closer to how they produced it at the pre-test.

Group 3: Bot, Bought, Boot. Figure 6.4 shows the normalised values for F1 and F2. As shown in the figure there were some changes in F1 values from post-test to the retention test. In particular, *bought* and *boot*, were produced with higher F1 values, in the HVPT and the HTP groups. In order to investigate any significant changes, separate linear mixed-effects models were built for F1 and F2 values.

For F1, the best fitting-model included time (pre, post, retention), training group (PT, HVPT, HT), and proficiency (HP, LP) coded as fixed factors, and participant and stimulus coded as random factors with random slope. The model excluded the interactions between time and proficiency, and the interaction between time, group and proficiency which means that these interactions are not significant for the analysis. The results from the model indicated a significant effect of time χ^2 (2) = 11.898, $p < .05$, suggesting a significant change in F1 values across time (pre, post, retention). The planned contrasts showed a significant change in F1 values from post-test to the retention test, b= 0.135, SE= 0.0402, pMCMC<.001, specifically learners produced *bought* and *boot* with higher F1 values in the HTP and HVPT groups. However, the change from pre-test to the post test was not significant.

There was no significant effect of training group, p>.05, but there was a significant two-way interaction between time and group, χ^2 (4) = 12.181, p<.05. The planned contrasts showed a significant change in F1 values from post-test to the retention test for HTP learners, b= -0.4482, SE= 0.169, pMCMC<.05. Specifically, learners produced *bought* and *boot* with a higher F1 at the retention test, compared to those in the PT group. However, there were no significant difference between the PT and the HVPT groups.

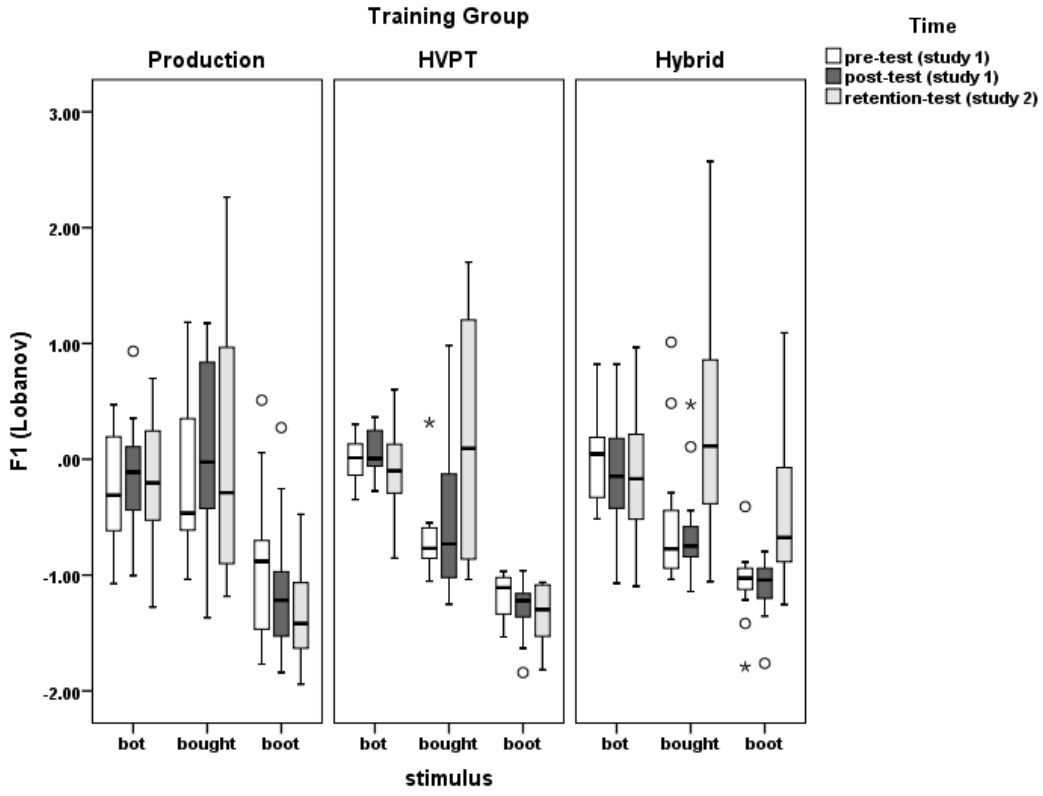


Figure 6.7: Boxplots showing F1 values for Vowel Group 3 (bot, bought, boot) produced by L2 learners in the 3 training conditions (PT, HVPT, HT). Formant values were normalised using Lobanov's method.

Additionally, although there was no significant main effect of proficiency level, there was a significant two-way interaction between training group and proficiency, $\chi^2(2)=11.52$, $p<.05$. The planned contrasts indicated that the performance of LP learners in the HTP group differed significantly from the LP learners in the PT group, b=-0.39, SE= 0.13, pMCMC<.05. Specifically, LP learners in the HTP group produced *bought* with higher F1 values than LP learners in the PT group (see Figure 6.8).

For the F2 values, the best fitting model included time (pre, post, retention), training group (PT, HVPT, HPT), and proficiency (HP, LP) coded as fixed factors, and participant and stimulus coded as random factors with random slope. The results from the model indicated no significant change in F2 values, which suggest no significant change in F2 values between groups and/or across time.

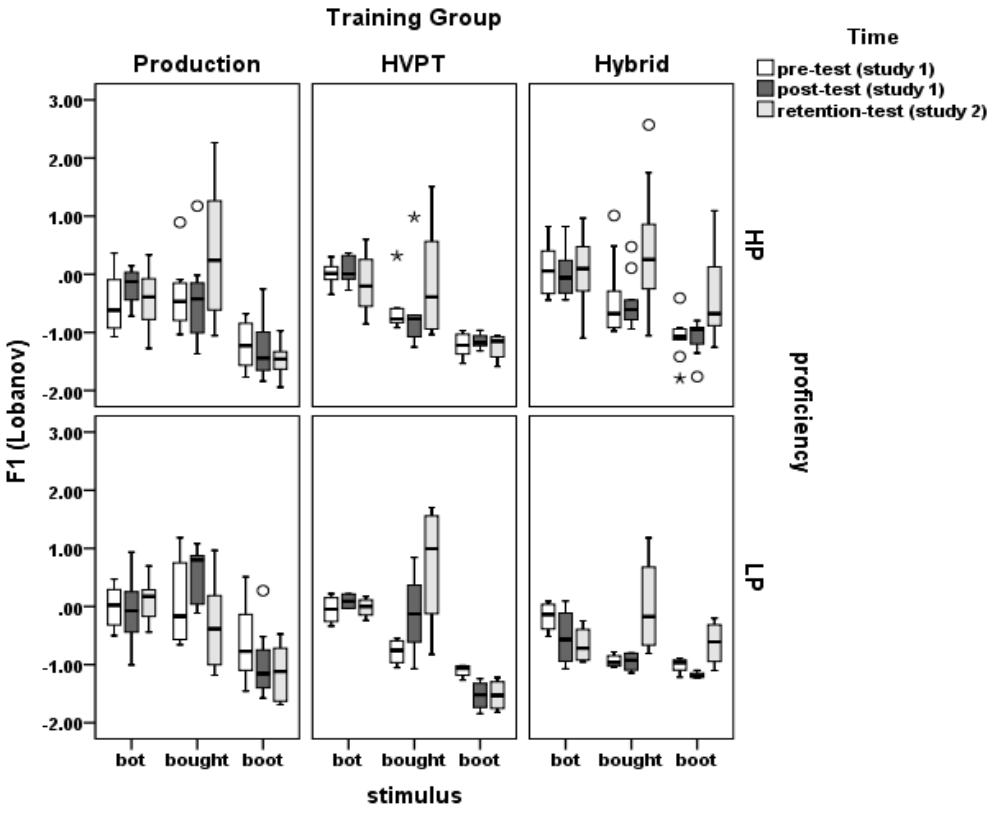


Figure 6.8 Boxplots showing F1 for Vowel Group 3 (bot, bought, boot) produced by L2 learners in the 3 training conditions (PT, HVPT, HT). Formant values were normalised using Lobanov's method, split by proficiency level HP=high proficiency, LP= Low proficiency (for the PT group; 5 LP & 3 HP; HVPT, 3 LP & 3HP; and for the HT group 6 HP & 2 LP).

6.3.3.2 Duration

Group 1: Beat, Bit, Bet, Bert. As displayed in Fig. 6.9 although all participants were correctly distinguishing long and short vowels, there appeared to be some changes in vowel duration for *beat*, *bert*, *bet* and *bit*, from pre- to post-test (Study 1), but no reliable changes from the post-test to the retention test.

In order to test for any significant change in duration across time (pre, post, retention) and across training groups (PT, HVPT, HTP), a linear mixed effects model was built for the duration data for this vowel group. As in Study 1 (Chapter 5), linear mixed models were built for the duration data based on the duration of the vowels (beat, bit, bet, and bert) in milliseconds (continuous scale). The best fitting-model was chosen using a top-down approach, which excludes ineffective factors after including all possible factors. The best fitting-model included time (pre, post, and retention), training group (PT, HVPT, HTP) and proficiency (HP, LP) coded as fixed factors, and participant and stimulus coded as random factors with random intercepts.

The results from the model indicated a significant main effect of time (pre, post, and retention), χ^2 (2) = 24.66, $p < .001$, suggesting a change in vowel duration from pre- to post-test, and from post-test to the retention test. As expected, the planned contrasts showed a significant change in the duration values from the pre- test to the post-test, b=-7.358, SE= 3.408, pMCMC<.05. Specifically, learners produced *beat* and *bert* with a longer duration at the post-test, such that they made greater difference between short and long vowels (see Figure 6.9).

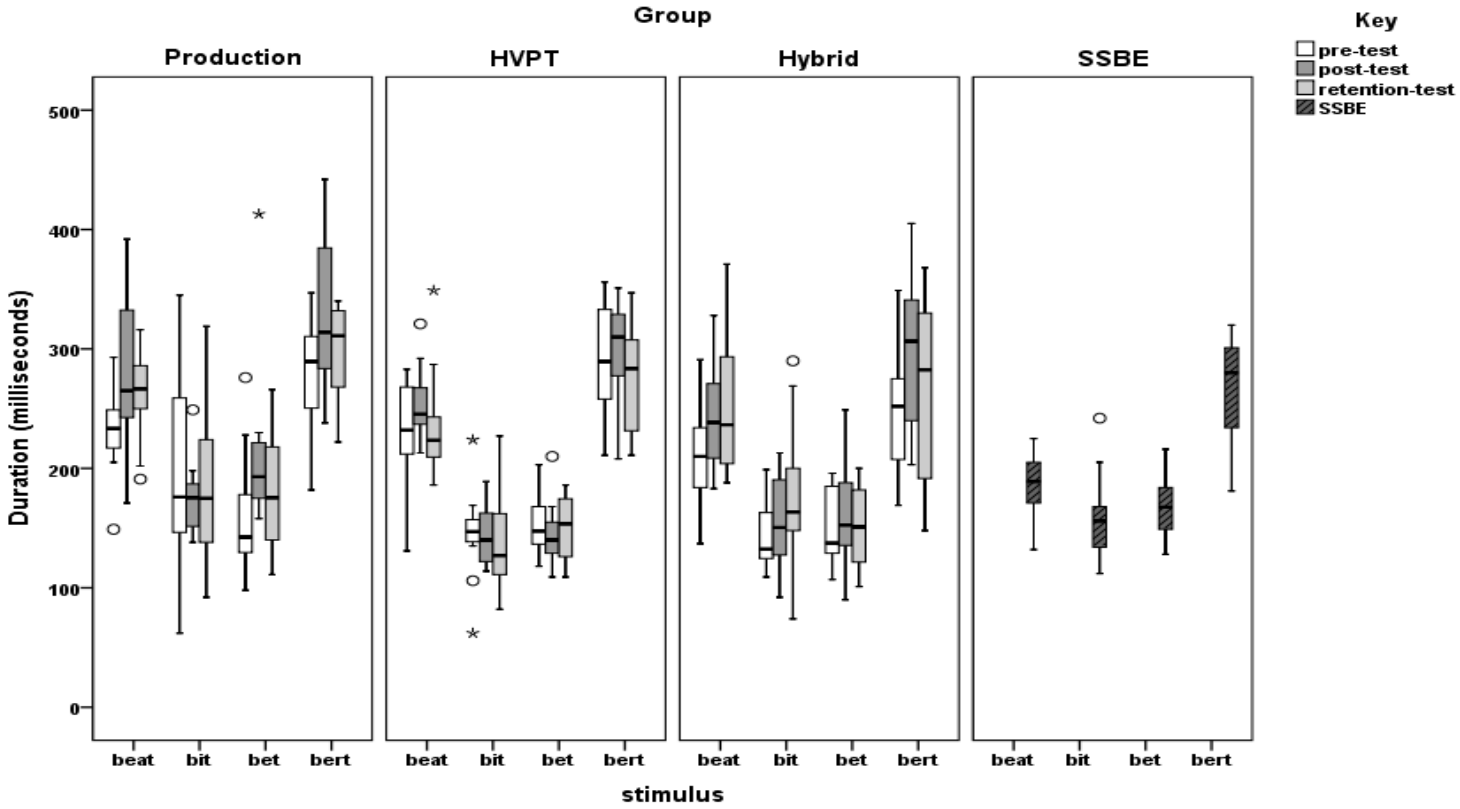


Figure 6.9: Boxplots showing vowel duration for Vowel Group 1 (beat, bit, bet, bert) produced by L2 learners in the 3 training conditions (PT, HVPT, HT), and compared to those of the SSBE speakers.

However, there was no significant change in vowel duration from post-test to retention-test, pMCMC>.05, indicating that any changes made from pre- to post-test had been retained. There was a significant effect of training group, χ^2 (2) = 11.70, p < .05, however, the planned contrasts did not show any significant differences between training groups, indicating that there were no reliable differences between training groups. There was no significant effect of proficiency, and no significant interaction between time and training.

Group 2: Bat, But, Bart. As displayed in Fig 6.10, there appeared to be some changes in vowel duration from pre- to post-test, and from post-test to the retention tests across training groups, though these were small, limited to particular vowels and particular groups (e.g., *but* in PT, and *bart* in HTP).

To investigate whether or not there were any significant changes in vowel duration in this group, mixed-effects linear models were built for the duration data for this vowel group (the best fitting model was chosen with the same top-down procedure). The best fitting-model included time (pre, post, retention), training group (PT, HVPT, HPT), and proficiency (HP, LP) as fixed factors, and participant and stimulus as random factors with random slopes.

The results from the model indicated a significant main effect of time, χ^2 (2) = 23.562, p < .001. The planned contrasts showed a significant change in vowel duration from pre- to post-test, b=-14.479, SE=3.889, pMCMC<.001. Learners produced *bart* with longer duration at the post-test, especially in the HTP group. However, there was no significant change from post- to the retention test, confirming that learners used similar vowel durations at the retention- test to those at the post-test.

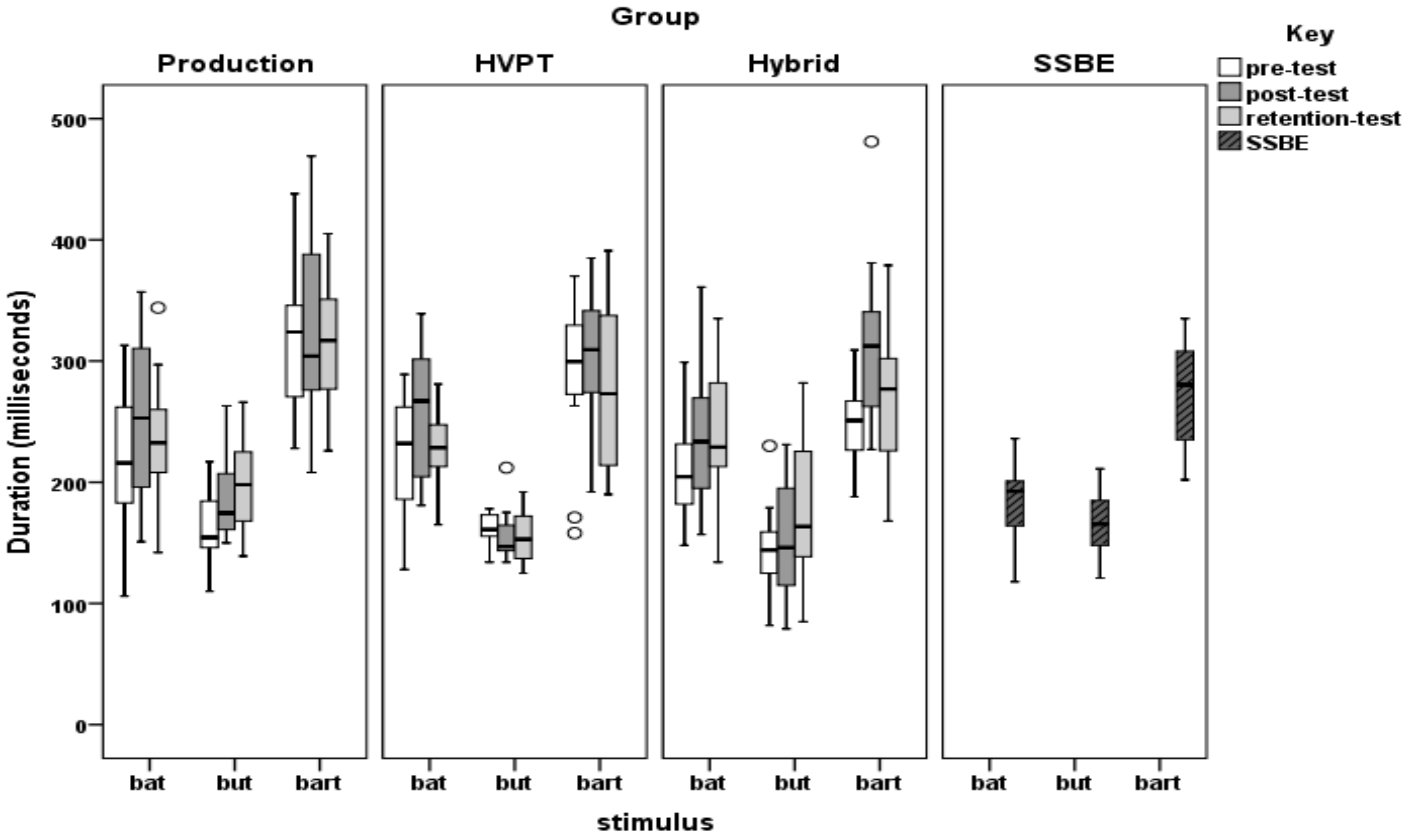


Figure 6.10: Boxplots showing vowel duration in milliseconds for vowel group 2 (bat, but, bart) produced by L2 learners in the 3 training conditions (PT, HVPT, HT), and compared to those of the SSBE speakers.

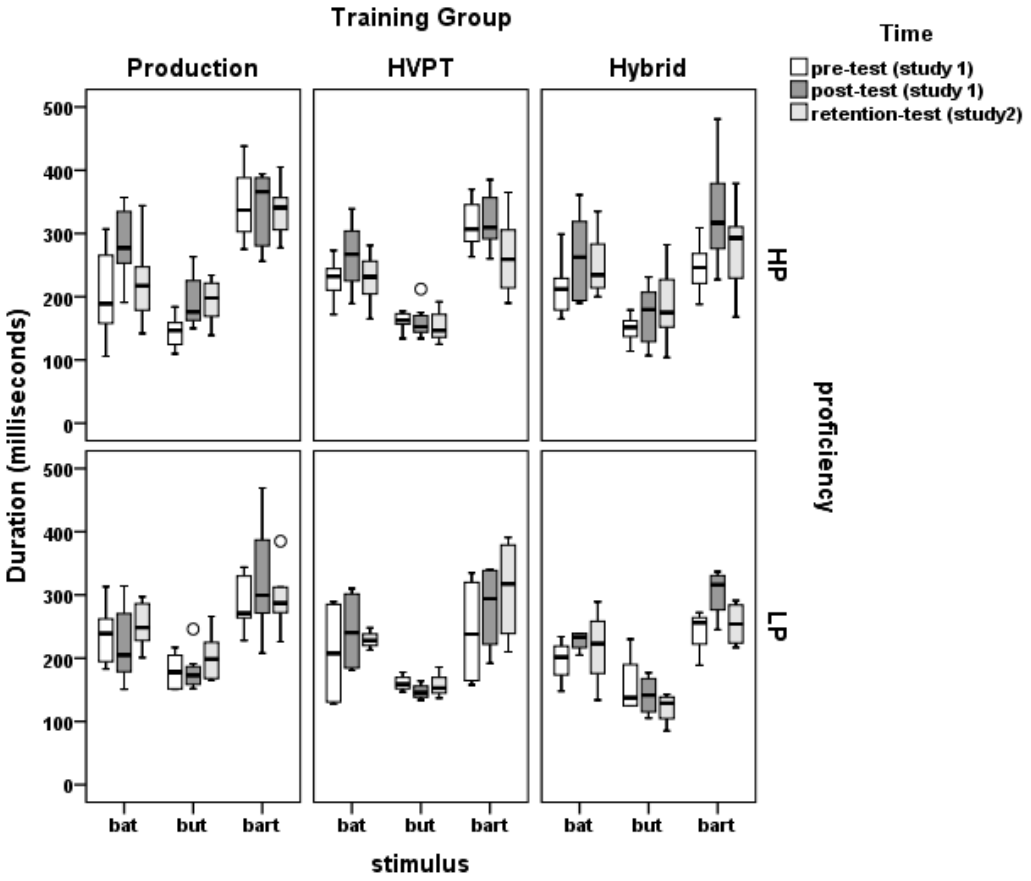


Figure 6.11: Boxplots showing vowel duration for vowel group 2 (bat, but, bart) produced by L2 learners in the 3 training conditions (PT, HVPT, HT), split by proficiency level (HP& LP= Low proficiency (for production group; 5 LP & 3 HP; HVPT, 3 LP & 3HP; and for the hybrid group 6 HP & 2 LP).

There was no significant effect of proficiency. However, there was a significant three-way interaction between time, training group, and proficiency, χ^2 (4) = 11.101, $p < .05$. The planned contrasts showed that there was a significant change in the vowel duration used by HP participants, but not the LP participants, in the PT group from pre- to the post-tests, compared to that of HP participants in the HVPT group, b= -22.362, SE= 9.79, pMCMC<05. At the retention test, HP learners in the PT group retained similar durations for *but* and *bart*, but changed the duration of *bat*, so that it was closer to their production at the pre-test. In contrast, HP participants in the HVPT

condition did not reliably change their vowel duration from post-test to retention test, b= 25.282, SE= 9.860, pMCMC<.05 (see Figure 6.11). The LP participants in both groups did not change their vowel duration from pre- to post- or from post- to the retention test.

Group 3: Bot, Bought, Boot. As displayed in Fig. 6.12 there were some changes in duration for this vowel group between different tests and training groups (e.g., the duration of *boot* from pre- to post- test in PT and HT groups).

To investigate if there were any changes in duration in this vowel group, a linear mixed-effects model was built for the data as described above. The best-fitting model included time (pre, post and retention) and training groups (PT, HVPT, HTP), and the interaction between time and group coded as fixed factors, and participant and stimulus coded as random factors with random slopes. The model excluded all other possible interactions (time, training group, and proficiency; time and proficiency; training group and proficiency) indicating that these interactions were not significant for the analysis.

The results from the model indicated a significant effect of time, χ^2 (2) = 41.417, p<.001, suggesting a change in the vowel duration from pre- to post-test and possibly from post-test to the retention test. The planned contrasts showed a significant vowel duration change from pre- to post-test, b= -20.77, SE= 3.650, pMCMC<.001, such that all participants tended to change the duration values of this vowel group after training, such that they produced these vowels with longer duration values after training (see Figure 6.12). However there was no significant change in the vowel duration from post to the retention test, pMCMC>.05.

The main effect of training group was significant, χ^2 (2) =20.092, p< .001, suggesting that participants in different training groups performed differently.

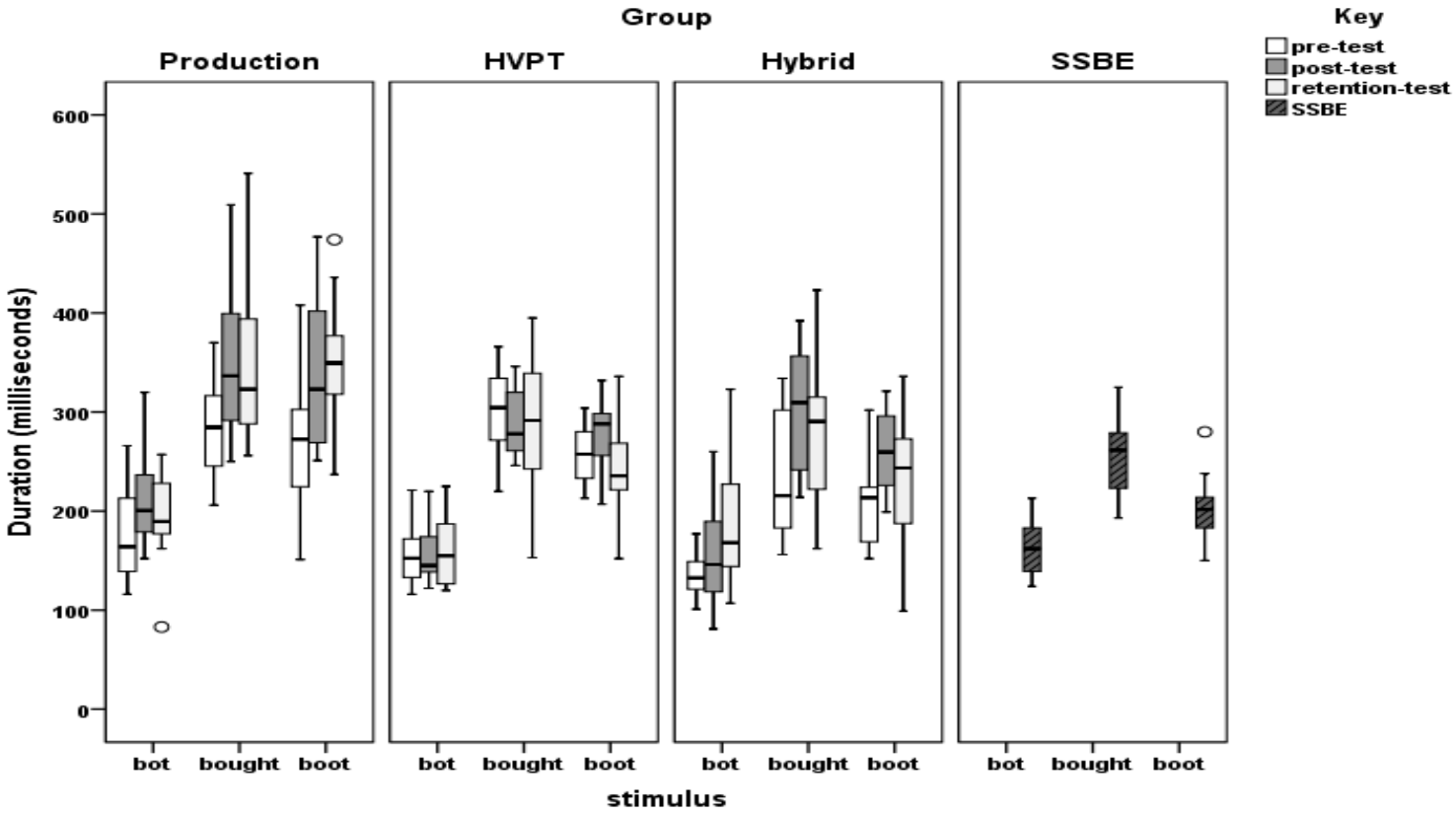


Figure 6.12: boxplots for vowel duration for vowel group 3 (bot, bought, boot) produced by L2 learners in the 3 training conditions (PT, HVPT, HT), and compared to those of the SSBE speakers.

The planned contrasts showed a significant difference in vowel duration for participants in the PT group compared to those in the HTP group, b= -23.239, SE=7.377, pMCMC<.05. Participants in the PT group tended to produce longer vowels. However, there was no significant difference between the PT and the HVPT groups.

There was a significant two way interaction between time and training group χ^2 (4) = 17.675, p< .001. The planned contrasts showed a significant change in the vowel duration from pre- to post-test for participants in the PT group compared to those in the HVPT group, b= 22.380, SE= 5.40, pMCMC<.001. Participants in the PT group produced longer vowels at the post test than did participants in the HVPT. There was also a significant difference in vowel duration from post-test to the retention test for participants in the PT condition compared to those in the HVPT group, b= -12.324, SE= 5.435, pMCMC<.05; participants in the PT group produced *bot* and *bought* with similar durations at the post-test and retention test, but further lengthened *boot*. HVPT learners also produced *bot* and *bought* with similar durations at the post-test and retention test, but shortened *boot*.

Summary. Study 1, demonstrated that there were changes in vowel production but that these were limited to a small number of vowels, specifically contrasts in vowel group 1 and vowel group 2. Specifically, the changes were in the F1 values for /ɪ/ and /e/ and F2 values for /ʌ/ and /ɑ/; in both cases, learners adjusted their formant frequency values to better match those of native speakers. These changes were limited to those learners who completed the PT and the HTP programs. For vowel duration participants in PT and HTP produced some vowels, *bert, bart, bought, boot*, longer after training, and produced *beat, bought* longer than that of native speakers.

Though these changes were limited, the results from this study showed that Learners who had completed the PT condition made the most changes to production as a result of training and retained this learning. Learners in the HTP programme retained similar vowel duration, whilst those in the HVPT condition made few changes to production initially and despite being resident in the UK, showed little evidence of further change.

In terms of spectral change, learners primarily changed *bit-bet* and *bought*. Learners in the PT group changed their production of these vowels to better match native speakers and retained these changes 6 months after training. However, learners in the HTP group who adjusted their production for this vowel contrast from the pre- to post- test, did not retain learning, producing this contrast more like they did at the pre-test. Additionally, PT and HTP learners produced *bought* with longer duration than that of the natives whilst HVPT and HTP learners produced *bought* with higher F1 values at the retention test.

For duration, all learners were able to make appropriate distinctions between short and long vowels at the pre-, post- and retention tests. However, some learners did appear to make reliable changes to their production as a result of training. All learners tended to produce long vowels, in particular *beat*, *boot*, and *bought*, with longer durations than native speakers. After training, learners in the PT group tended to increase the length of long vowels further and retained this learning at the retention test.

6.4 Discussion

The main aim of this study was to investigate whether any of the changes in vowel perception and production that occurred as a result of the training given in Study 1, were retained 6 months after training. Also of interest, was whether or not retention was affected by training type; PT (production-based), HVPT (perception-based), or HTP (production and perception).

Previous studies have shown that changes to vowel identification as a result of HVPT are retained over relatively long periods of time (e.g., Bradlow 1999, Iverson & Evans, 2009). Similar findings have been found in this study: participants who took part in the HVPT or HTP training programmes improved initially in their performance on the vowel identification task, and retained these improvements 6 months after they had completed training. Additionally, in the current study, participants further improved in their performance on this task, such that their performance was better at the retention test than at the post-test (Study 1).

The results for the PT training group were more complex. Study 1 showed that learners in the HVPT condition improved more in their vowel identification than those who had completed PT. Only a subset of participants were tested at the retention test and in this analysis, which investigated improvements during training and compared these to the results from the retention test, there was no significant interaction of group and time. This suggests that participants in the PT condition improved in their vowel identification as much as those in other training conditions, and that at least for some learners, learning is not as domain-specific as the results from the training study might have initially suggested.

Additionally, HP learners in the PT training condition also performed more poorly overall than did HP learners in the HVPT and HTP training conditions, such that this subset of participants were not as well balanced in terms of proficiency. This suggests that performance on the grammar test that was used to assign participants to proficiency groups may not always predict performance with spoken language. However, although there was no significant interaction between time, training group and proficiency, observation of the data (Figure 6.2) indicated that improvements to vowel identification in the PT training condition were primarily for LP learners and that these learners improved more than HP learners in the HVPT and HTP conditions from the post-test to retention test, such that at the retention test performance was similar across the different proficiency levels and training groups. One reason for this could be that production training served as a key for learning. Given that it is impossible to complete production training without involving perception, it is likely that learners had acquired some perceptual knowledge during the PT sessions, and that with time and further exposure to English, had adjusted their vowel category perception. Thus, although training itself might be initially domain-specific, the knowledge learners gain may enable them to later develop their skills in another domain.

Study 1 demonstrated that LP learners improved more than HP learners in HVPT and HTP, possibly because they have more scope for learning. In this study, though there was no main effect of proficiency, there was an interaction between time and proficiency. HP learners improved most from the pre- to post-test and retained this

learning, with those who had received production training (HPT and PT training groups) improving further. One possibility is that production training helped L2 learners in the long-term, such that this type of training, involving both articulatory and perceptual skills, served as a key for deeper learning. In contrast, although HVPT may have initially enabled learners to identify sounds more accurately, without continued reinforcement, this learning may not have had such long-lasting effects.

In contrast, all LP learners, regardless of training modality, showed more improvement in speech recognition in noise from the post-test to the retention-test than from pre- to post-test than did HP learners. This is likely because the LP learners started from a lower level, and thus had more room for improvement. However, all learners, except for HP learners in the HVPT condition, showed further improvement from the post-test to retention test. Participants in this study were enrolled as students at universities in London. It is possible that when they initially received training (Study 1), they used it as base to build more perceptual knowledge during their daily interactions with native speakers. They would have been likely to have had more exposure to English through their studies after they had completed the training, and this in turn, may have led to further improvement in their lexical knowledge 6 months after training. A combination of improvements in perceptual and lexical knowledge may thus have enabled them to improve further in their performance in the sentence recognition in noise task, a task in which lexical knowledge may play an important role in resolving ambiguity (see e.g., Mattys et al., 2012).

As in Study 1, participants who completed HVPT showed no change in production from pre- to post test, and as expected, there were no further changes from the post-test to retention test. However, participants in the PT group who had changed their production from the pre- to post test, retained their vowel production especially for /ɪ/-/e/ contrasts. Participants in the HTP group also changed their production of the /ɪ/-/e/ contrast from pre- to post-test, but they did not retain this learning. One possibility for this pattern of results is the single production-based training session that these participants completed, which was enough to modify their vowel production, at least for this vowel contrast in the short-term, did not lead to long-term changes. This suggests that although a small amount of production training can effect changes in

pronunciation, in order for long-term modification to pronunciation to take place, a relatively large amount of training is needed.

In summary, the current results suggest that not only HVPT but also production-based training yields long-term learning effects. The subset of learners in these PT, HVPT and HTP training groups showed learning in vowel identification and speech in noise after training and this learning was retained 6 months later, with some evidence for further learning from the post-test to retention test. However, for production, it seems that only those in the PT group retained improvements in vowel production. Overall these results suggest that though production-based training might initially be domain-specific, in particular for LP learners, combining production and perceptual learning might have long-lasting advantages for second language learning.

Chapter 7 Training Arabic learners of English on vowel production: comparing the efficiency of production training in an immersion and non-immersion settings.

7.1 Overview

In previous chapters it was argued that L2 learners seem to benefit from different types of training, and that dependent on the training received, they retain improvements in their perception and/or production abilities at least 6 months after training. As in many L2 studies, the training and retention studies presented here were conducted in the UK where students were not only getting training in the lab, but also getting training through continued experience of interaction with native speakers. This makes it difficult to know exactly what influence the training is having vis-à-vis the environment. This study aimed to investigate the efficacy of these kinds of training techniques in non-immersion settings, specifically, whether or not the benefits of production training might differ in a non-immersion setting, where experience of interacting with native speakers is much harder to find.

7.2 Introduction

Auditory phonetic training has been proven to be highly successful in improving learning of difficult L2 phonemes. Most of these studies have used HVPT where listeners listen to and identify phonemes produced in different contexts by multiple speakers, and receive corrective feedback on their responses (e.g., Bradlow et al, 1997; Lively et al, 1992, Iverson & Evans, 2009; Nishi & Kewley-Port, 2007). Though some studies have trained learners on entirely new phonemic contrasts in a language that they do not use (e.g., Hirata, 2004; Pruitt et al., 2006), many have focused on English (i.e., training L2 learners of English on English phonemes) and have been with L2 learners living in an English-speaking country (e.g., Iverson & Evans, 2009; Hattori, 2009).

More recently Iverson et al (2012) used HVPT to train French learners of English on English vowels comparing two groups with English experience; French speakers in France (inexperienced learners), and French speakers in London (experienced learners), Despite the fact that the French speakers in London had many

more opportunities to interact with native English speakers, the results demonstrated that both training groups improved similarly. However, the production training studies, and the studies that used CALL for articulatory training (see Chapter 4, p. 92 for review), have not, at least to my knowledge, compared training in the learners' L1 environment, where there are few opportunities to interact with native speakers and reinforce training, with training in the L2 speaking country.

The present study aims to compare the potential benefits of production training for learners in two different settings; 1) training in the L2-speaking country (immersion setting) and 2) training in the home country (non-immersion setting). One possibility is that speakers who are trained in the L2-speaking country and who have opportunities to consolidate learning through daily interaction with native speakers, may improve more than those trained in the L2 speaking country, where they do not regularly interact with native speakers. On the other hand, production training might be more successful for learners who have less exposure and fewer interactions with native speakers. This is because the production training used here emphasizes exposure to natural stimuli produced by native speakers. Those who live in an English speaking country and have daily interactions in their L2 are already exposed to richer array of stimuli than can be delivered by several sessions of training, and thus, they may receive little additional benefit from this type of focussed training.

Ten L1 Arabic speakers, living in Jeddah, Saudi Arabia, were given five one-to-one sessions of production training on British English vowels using CALVin (see p. 106). To investigate potential changes in perception and production, they completed the battery of pre- and post-tests used in the UK-based training study (see Chapter 5). To investigate whether or not any improvements in perception and production were affected by learning environment, the results were compared with those of participants who had completed production training in London (Chapter 5).

7.3 Methods

7.3.1 Participants:

Ten participants took part in this study, but only 9 (4 male) completed all the sessions; 1 participant failed to complete the post-test. Participants were aged 19- 43

years old (median 34 years old), and had begun learning English when they were 3-19 years old (median 13 years old). Only one participant reported having lived in an English speaking country, and had lived in UK for one year, a long time ago (10 years). Participants reported no regular interactions with native speakers at the time of testing. Although 3 participants were recruited from a language centre where some of the teachers are native speakers of English, at time of testing they were not taught by the native speaker teachers. All participants completed the written grammar section of the Oxford placement test (Allan, 1992) to evaluate their English proficiency.

As in the training study reported in Chapter 5, 10 Standard Southern British English speakers (4 males) participated in the study. They were 18-40 years old (median 21 years old), recruited from UCL Psychology pool, and all were from the south of England. These participants rated Arabic learners' production for accent and intelligibility, and recorded the same /b/-V-/t/ words recorded produced by Arabic learners to give normative data.

7.3.2 Stimuli and apparatus

A. Pre- and post-tests

The stimuli were the same as in Chapter 5 (see p. 102)

B. Training

The same as in Chapter 5 (see p. 103)

7.3.3 Procedure

Pre/post tests were the same as in Chapter 5. The training protocol was also the same as for the production training in Chapter 5 (see p. 104).

7.4 Results

The results were compared with those of the PT group in Chapter 5. Participants in both Chapter 5 and this chapter were not deliberately recruited to be matched for proficiency, though their performance at pre-test showed that they performed similarly before training.

7.4.1 Perceptual tasks

7.4.1.1 Vowel identification

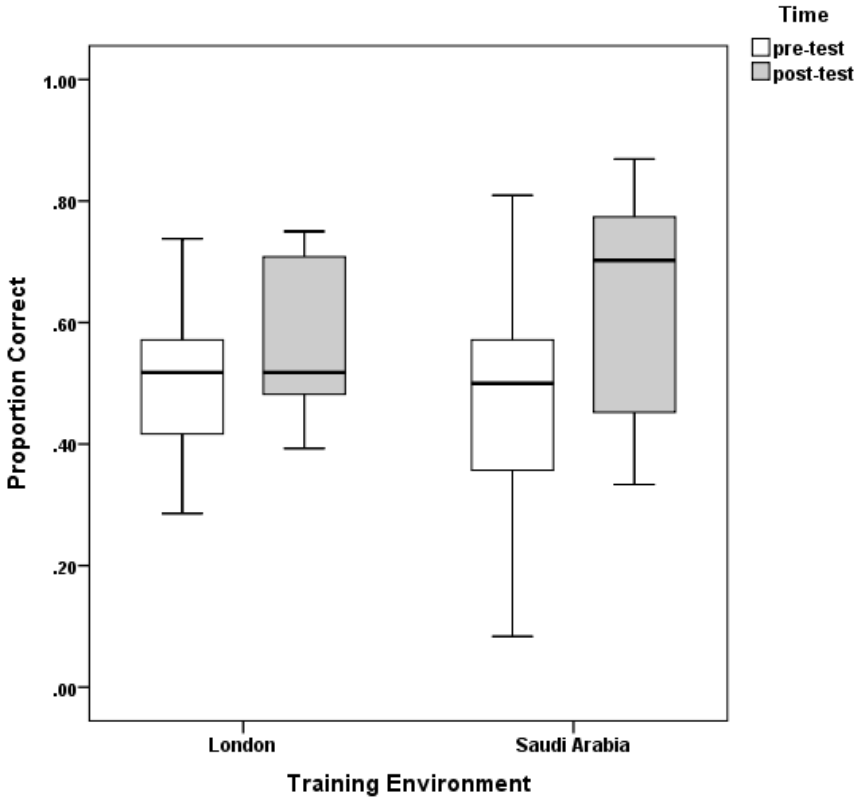


Figure 7.1: Boxplots showing overall performance (average proportion correct) on the vowel identification task across the two training groups; production training group in London (N=16), and production training group in Saudi Arabia (N=9).

Figure 7.1 displays the overall vowel identification accuracy for Arabic learners of English in the same training type condition (PT), in different immersion settings; one in London where subjects have the opportunity to regularly interact with native speakers, and the other where they were trained in Saudi Arabia (SA) in a non-immersion setting. The boxplots indicate that, the group that was trained in SA seemed to improve more from pre- to post-test than the group that was trained in London. This observation was tested by fitting a logistic mixed effects model for the identification binomial responses (i.e., correct/incorrect) for each vowel. The best-fitting model to the data was chosen with a top-down procedure (see Chapter 5, p. 115), and was fit by the Laplace approximation with time (pre, and post), and training environment (PT in London vs. PT in Saudi Arabia) coded as fixed factors, and participant and stimulus

coded as random factors with the random slope of time. The best fitting model excluded proficiency and the interactions between training environment and proficiency, time and proficiency, and the three way interaction between training environment, time and proficiency, which indicates that these factors are not significant. The random factors were added with random slopes for pre/post testing, so that the difference in the pre- and post-tests could be calculated for each stimulus, and for each participant.

The results from the model indicated that there was no significant effect of training environment, χ^2 (1) = 0.3268, p>.05. The main effect of time (pre and post) was significant χ^2 (1) = 35.65, p<0.001, indicating that there was a change in the vowel identification from the pre- to post-test. The planned contrasts confirmed that, there was an improvement in vowel identification from pre- to post-test, b=-0.276589, SE=0.046319, z=-5.971, p<.001, confirming that all participants improved in their performance from pre- to post-test.

However, there was also a significant two-way interaction between training environment and time, χ^2 (1) = 15.556, p<0.001 which indicate that subjects in one of the settings had changed their performance from pre- to post-test more than the other. The planned contrasts for the interaction between time (pre and post), and training environment (SA and London) demonstrated that the group that was trained in SA improved significantly more from the pre-to- post- test than the equivalent group in London, b=-0.16874, SE= 0.042784, z=-3.944, p<.001.

7.4.1.2 Category discrimination

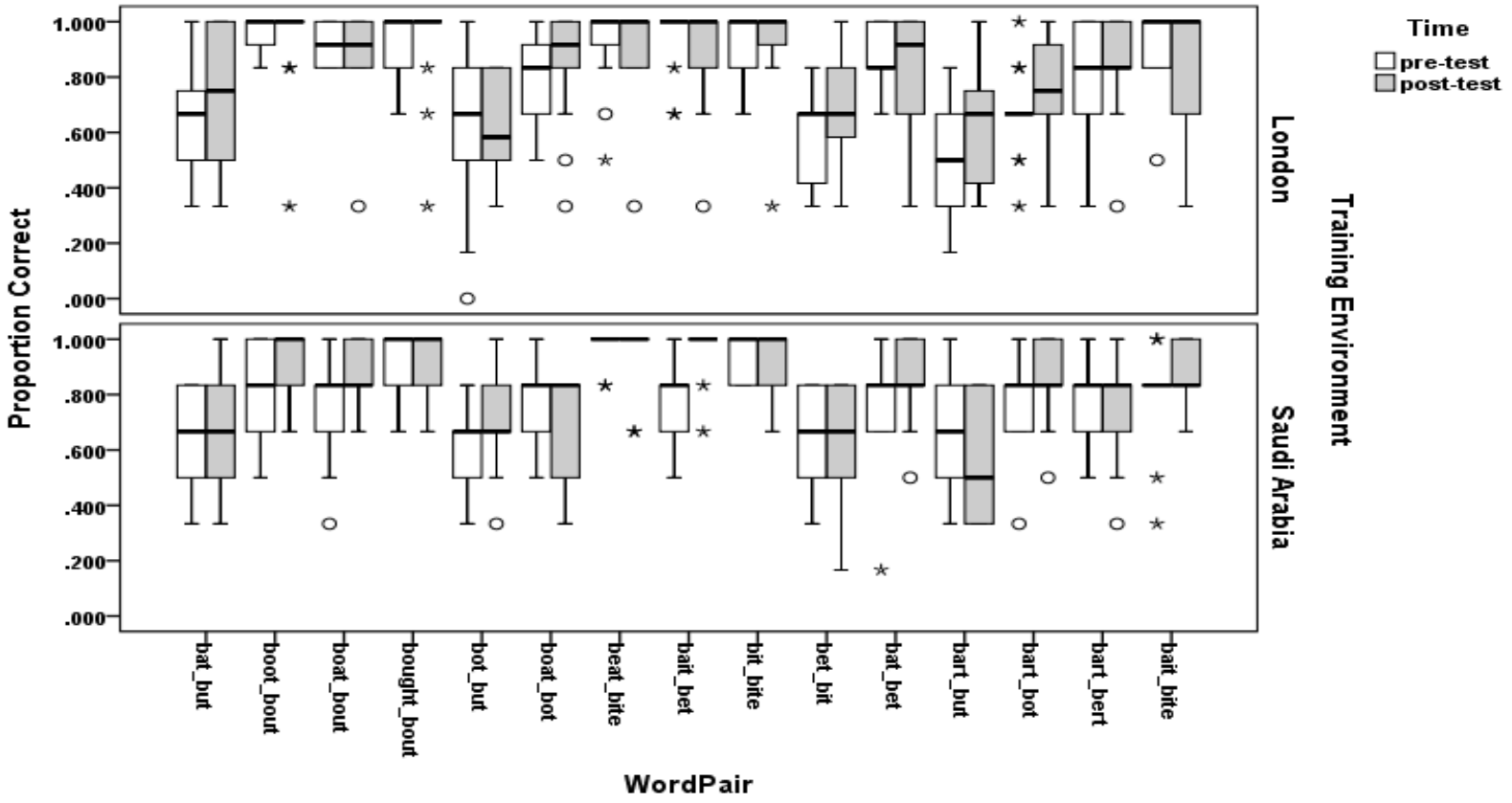


Figure 7.2: Boxplots showing performance on the category discrimination task for production groups in different environment (London vs SA). The y-axis shows the word- pair, and the x-axis shows the proportion correct. Participants who were trained in London are shown in the upper row (N=16) and those who were trained in Saudi Arabia in the lower row (N=9).

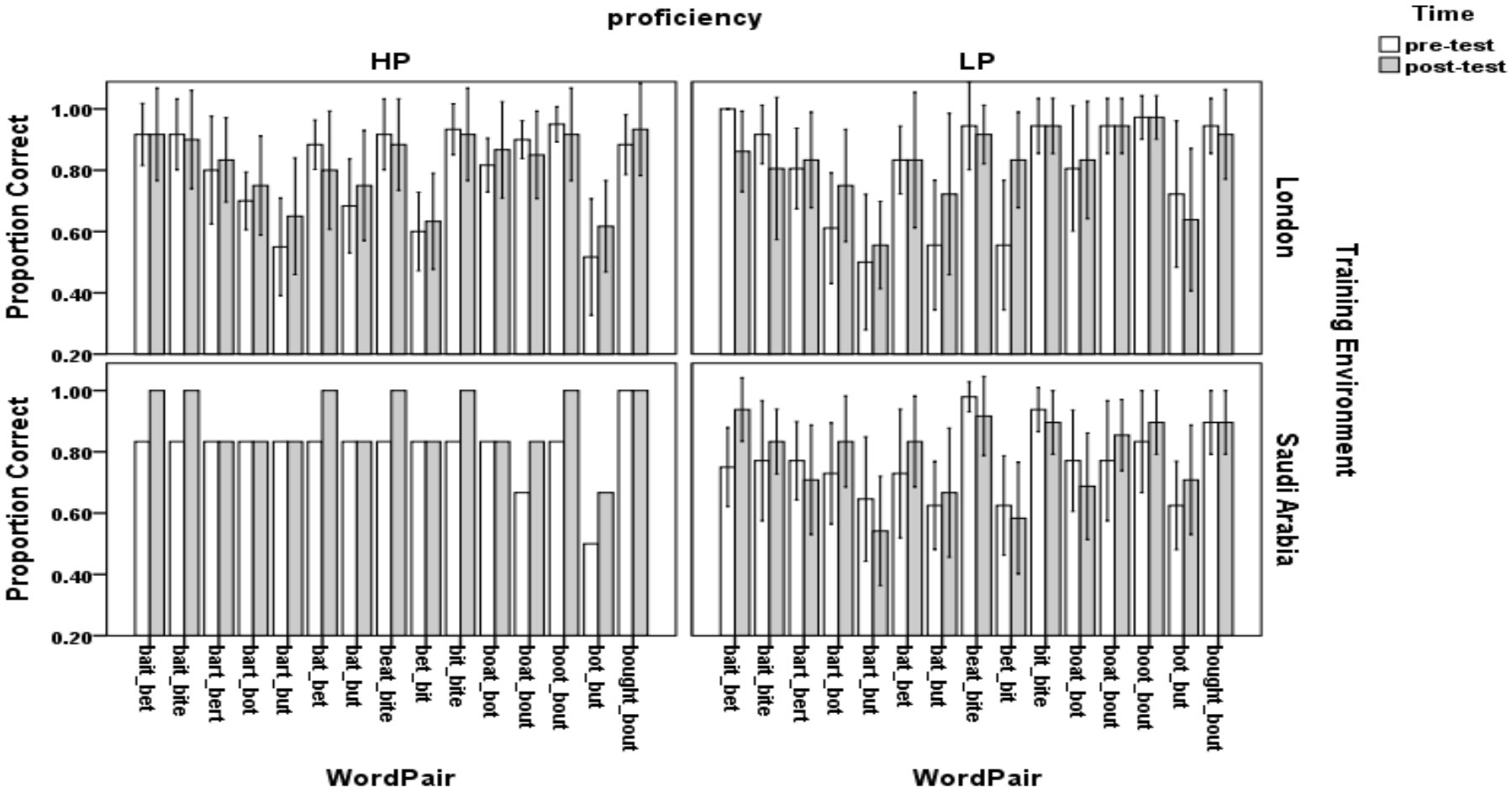


Figure 7.3: Bar chart showing the proportion correct for the category discrimination task in the two groups; in London (N=16) and in Saudi Arabia (N=9) in the rows, divided by the proficiency level in the columns (production in London, HP=10 participants, LP=6; production in Saudi Arabia, HP=1, LP=8).

Figure 7.2 displays the results the category discrimination task (London and in SA) in proportion correct. The average of the proportion correct of same word pairs was calculated (e.g., the average of *bat-bot*, and *bot-bat*, was calculated and merged into one word pair, *bat-bot*). It appears from the boxplots that there was no difference in word pair discrimination at pre- and post-test. In order to look for any changes, a linear mixed model was built for the data with up-down procedure (see chapter 5). The best fitting model included training environment (London vs. Saudi Arabia), and proficiency (HP, LP) as fixed factors and word pair as random factors with random intercept. There was no significant effect of the factors, which confirms that there was no overall significant change in category discrimination performance from pre- to post-test. However, there was a significant interaction between training environment and proficiency level, $\chi 2\ (1) = 6.866$, $p<.05$. The planned contrasts showed that the HP learners who were trained in London improved more than those trained in SA, b= -3.518, SE= 1.663, pMCMC<.05. However, this interaction may be driven by the single HP participant in SA group (see Figure 7.3), and thus, it is difficult to know how generalizable this result would be to a larger population.

7.4.1.3 Speech recognition in noise IEEE

As shown in Figure 7.4 participants who were trained in Saudi Arabia started off with a higher SNR (mean at pre-test = 16.7 dB), than those who were trained in London (mean at pre-test = 12 dB). After training, performance on this task appeared to improve more for those trained in SA (mean at post-test =5.9) than those trained in London (mean at post-test =9.4). In order to test these observations, a linear mixed-effects model was built for the data. The best fitting-model for the data included time (pre, post), training environment (London, Saudi Arabia), and proficiency (HP, LP) as fixed factors, and participant as a random with a random intercept. The main effect of time was significant, $\chi^2\ (1) = 6.661$, $p<.05$, suggesting a change in the performance from pre- to post-test. The planned contrasts indicated a significant change in the performance from pre- to post-test, b= 2.634, SE= 1.0205, pMCMC<.05, confirming that all subjects improved in their performance after training.

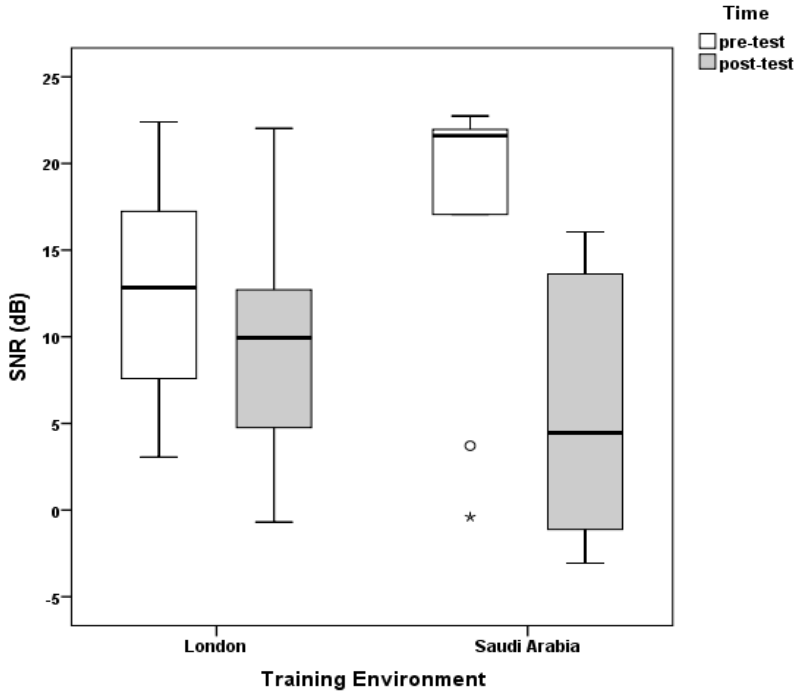


Figure 7.5: Boxplots of speech reception threshold (dB SPL) for L2 listeners across training environment (London N=16 & Saudi Arabia, N=9) at the pre- and post-tests.

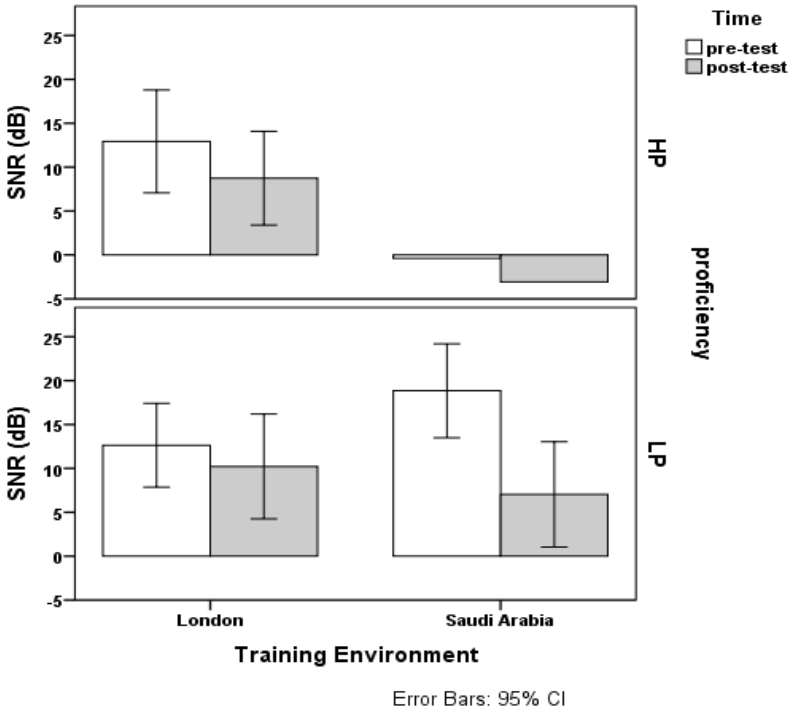


Figure 7.4: Bar chart of speech reception threshold (dB SPL) for L2 listeners across training groups at the pre and post-tests and across proficiency levels; High Proficiency (HP; London=10, Saudi Arabia=1), and Low Proficiency (LP; London =6, LP=8).

The main effect of training environment was not significant. However, the main effect of proficiency was significant, χ^2 (1) = 5.267, p<.05, suggesting that proficiency level affected listeners' performance. The planned contrasts indicated that the LP participants performed better at the post-test than HP ones, b= -3.8173, SE= 1.663, pMCMC<.05. There was also a significant two-way interaction between training environment and proficiency, $\chi 2$ (1) = 4.475, p<.05, indicating that proficiency affected performance differently in the two training environments. The planned contrasts showed that the LP participants who were trained in SA, performed better at the post-test compared to the equivalent proficiency group who were trained in London, b= -3.518, SE= 1.66, pMCMC<.05. However, there was no significant interaction between HP proficiency and training environment, possibly because the HP group in SA only contains one participant (see Fig 7.5).

7.4.2 Speech production

7.4.2.1 Acoustic analysis of /b/-V-/t/ words

7.4.2.1.1 Spectral Analysis

As in Chapter 5, in order to avoid multiple comparisons, the monophthongs were divided into three groups; Group 1 (beat, bit, bet, bert), Group 2 (bat, but, bart), and Group 3 (bot, bought, boot). Analysis of F1 & F2 for each vowel group will be presented first, then duration. As before, the formants values were normalised using Lobanov's method to enable data from male and female participants to be compared.

Group 1: Beat, Bit, Bet, Bert. As displayed in Figure 7.6, there was little evidence of change in F1 for Group 1 after training. In order to investigate any spectral changes for this group of vowels (*beat, bert, bet, bit*) after training, separate linear mixed-effects models were built for F1 and F2.

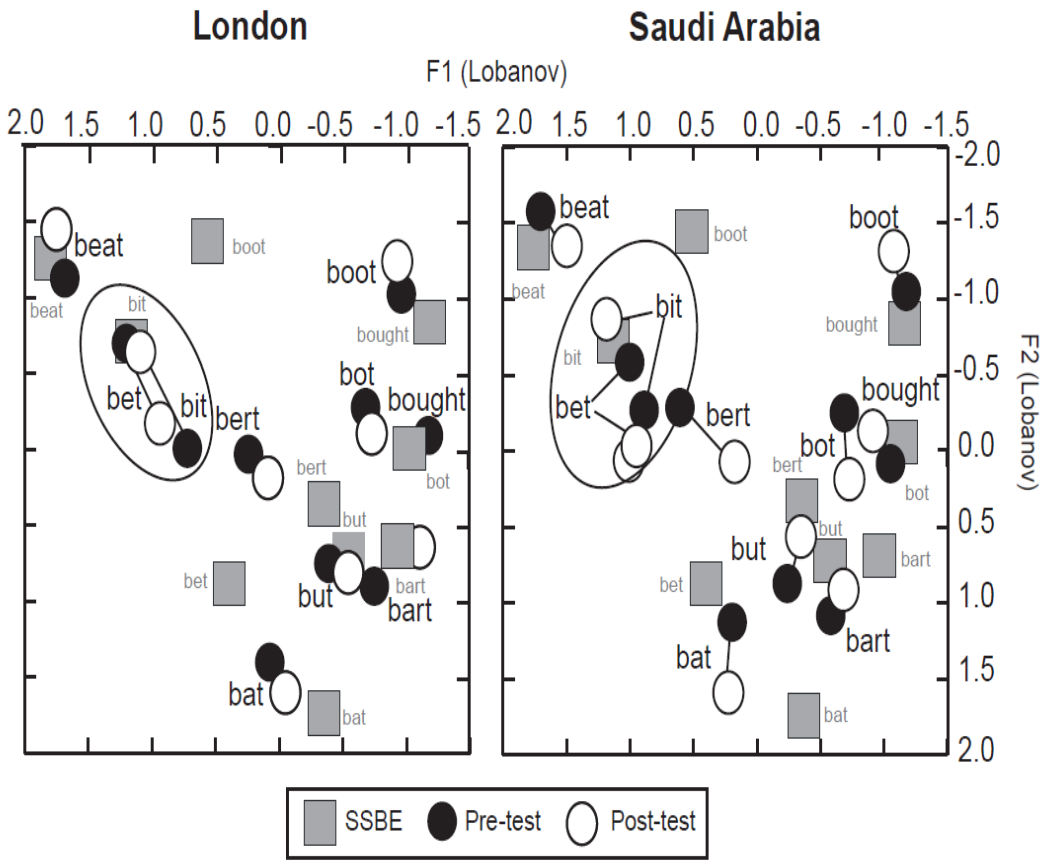


Figure 7.6: Average F1 and F2 formant frequency plots for London and SA subjects' productions of target words. Productions from the pre-test (dark circles) and post-test (white circles) are plotted with measurements from SSBE speakers (grey circles).

The best fitting-model for F1 included time (pre-post) and training environment (London & SA). The main effect of time was not significant, suggesting that there was no significant change in F1 values from pre- to post-test. The main effect of training environment was not significant, indicating that there was no significant difference between groups in London and SA in changing the F1 values.

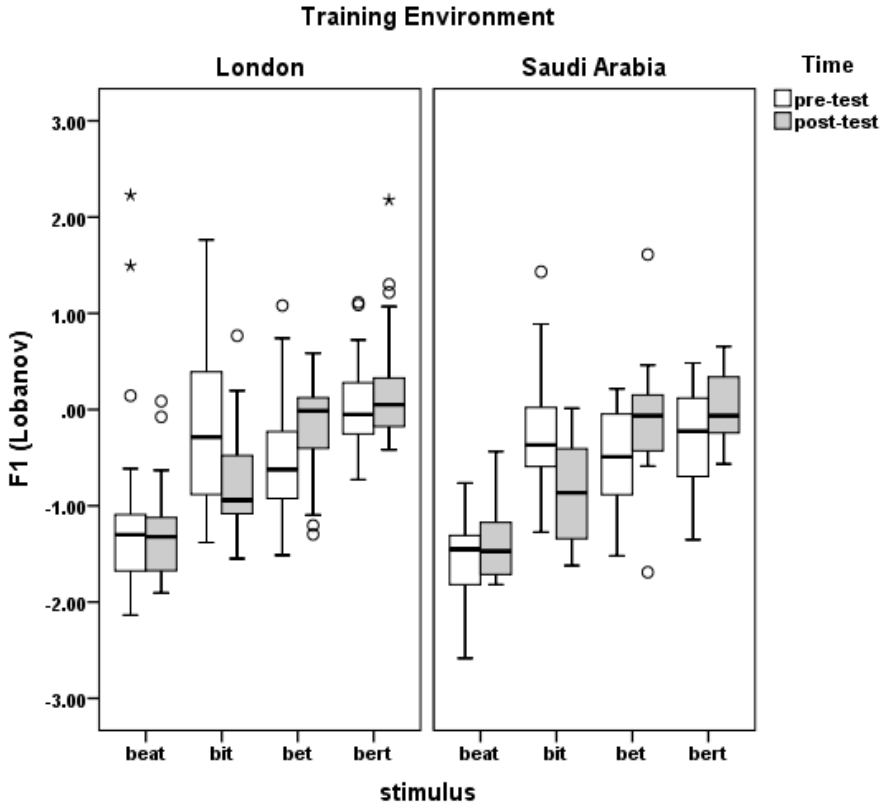


Figure 7.7: Boxplots showing F2 values for vowel group 1 (*beat, bert, bet, bit)* produced by L2 learners in the two training environments (London, N=16, Saudi Arabia, N=9) at the pre- and post-tests. The F2 values for stimuli was the average of 2 repetition of a word for each speaker.

This was surprising as in both the boxplots and vowel plot (see Fig 7.6 & Fig 7.7), learners in both group environments appeared to alter F1 values for /ɪ/-/e/, though the change appeared to be smaller for those who were trained in Saudi Arabia. At the pre-test learners produced the vowel /ɪ/ with higher F1 values, and the vowel /e/ with lower F1 values. However, after training they altered their F1 values for this contrast such that they produced /ɪ/ with lower F1 values, and /e/ with higher F1 values, so that these vowels were more similar to native F1 values for this vowel contrast. Participants also produced a more central vowel for (*bert)* (see Fig 7.6).

For F2 values, the best fit model included time (pre-post) and training environment (London & SA) as fixed factors, and stimulus and participant as random

factors with random slopes of time. There was no significant effect of the factors, indicating no significant change in F2 values from pre- to post-test.

Group 2: Bat, But, Bert. As displayed in Figure 7.6, there appears to be a small change from pre- to post-test in F2 values, but not in F1 values. In order to investigate any potential spectral changes for this group of vowels (*bat, but, bart*) after training, a linear mixed-effects model was built for the normalized data for F1 and F2 separately.

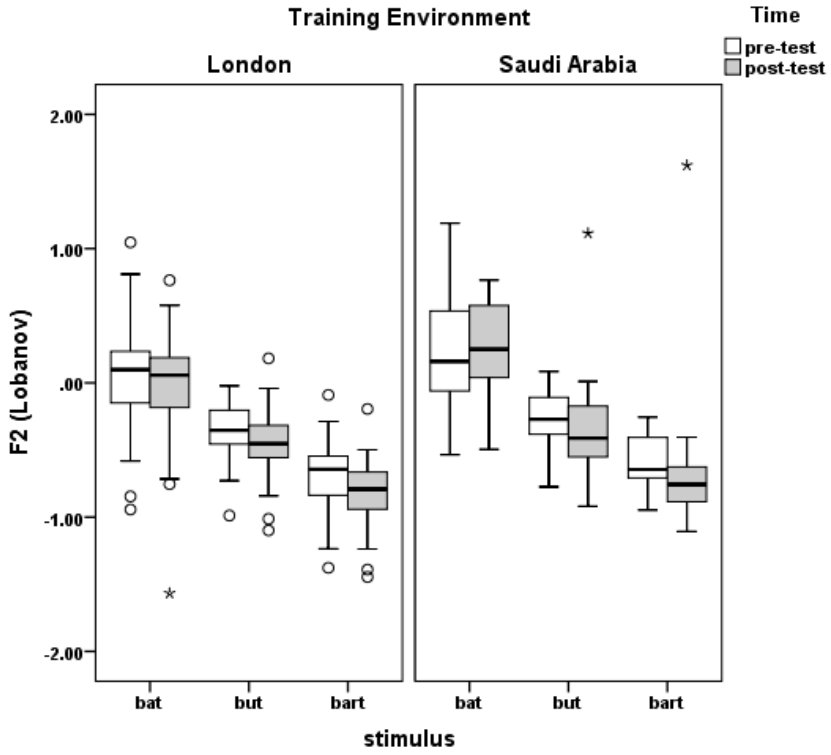


Figure 7.8: Boxplots showing F2 values for vowel group 2 (*bat, but, bart)* produced by L2 learners in the two training environments (London, N=16, Saudi Arabia, N=9) at the pre- and post-tests. The F2 values for stimuli was the average of 2 repetitions of a word for each speaker.

The best fitting-model for F1 included time (pre-post), training environment (London vs. SA), and proficiency (HP & LP) as fixed factors, and stimulus and participant as random factors with random intercepts. There was no significant effect of the factors suggesting no significant change in F1 values. For the F2 values, the best fitting model included time (pre, post) training environment (London, SA) and proficiency (HP, LP) as fixed factors, and participant and stimulus as random factors with random intercepts. The main effect of time was not significant, which means that there was no change in F2 values after training. The main effect of training

environment was significant, χ^2 (1) = 6.770, p<.05, which suggests that the F2 values were different in different training environments. The planned contrasts indicated a significant difference in F2 values in the vowels produced by the group that was trained in Saudi Arabia compared to those produced by equivalent group in London, b=-0.088, SE=0.029, pMCMC<.05, especially for *bart* and *but* (see Fig 7.8). Subjects in SA tended to use lower F2 values than those in London, though these differences were very small and, as displayed in Fig 7.6, it is unclear if this result represents any reliable difference in the production of these vowels between the two groups in this dimension.

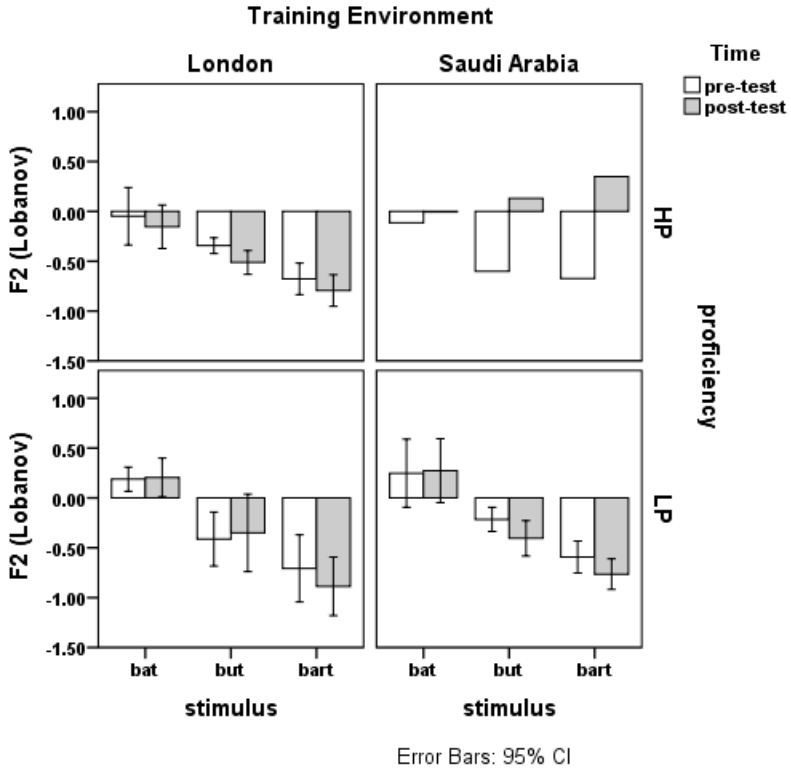


Figure 7.9: Boxplots showing F2 values for vowel group 2 (*bat, but, bart)* produced by L2 learners in the two training environments divided by proficiency levels [London (N=16, HP=10 participants, LP=6] and SA [N=9, HP=1, LP=8] at the pre- and post-tests.

There also was a significant three-way interaction between time, training environment and proficiency, χ^2 (1) = 12.733, p<.001. The planned contrasts indicated that the HP participants that were trained in Saudi Arabia changed their F2 values for this group of vowels after training more than those produced by HP participants who were trained in London, b= 0.1036, SE= 0.0290, pMCMC<.001(see Fig 7.9).

However, note that there was only one HP speaker in the SA group, and so it is hard to know how generalizable this finding is.

Group 3: Bot, Bought, Boot. As displayed in Fig 7.6, there seem to be few changes from pre- to post-test in these vowels in the F1 or F2 dimension. In order to investigate any potential changes in F1 and F2 values after training for this vowel group (bot, bought, boot), a linear mixed-effects model was built for F1, and F2 separately.

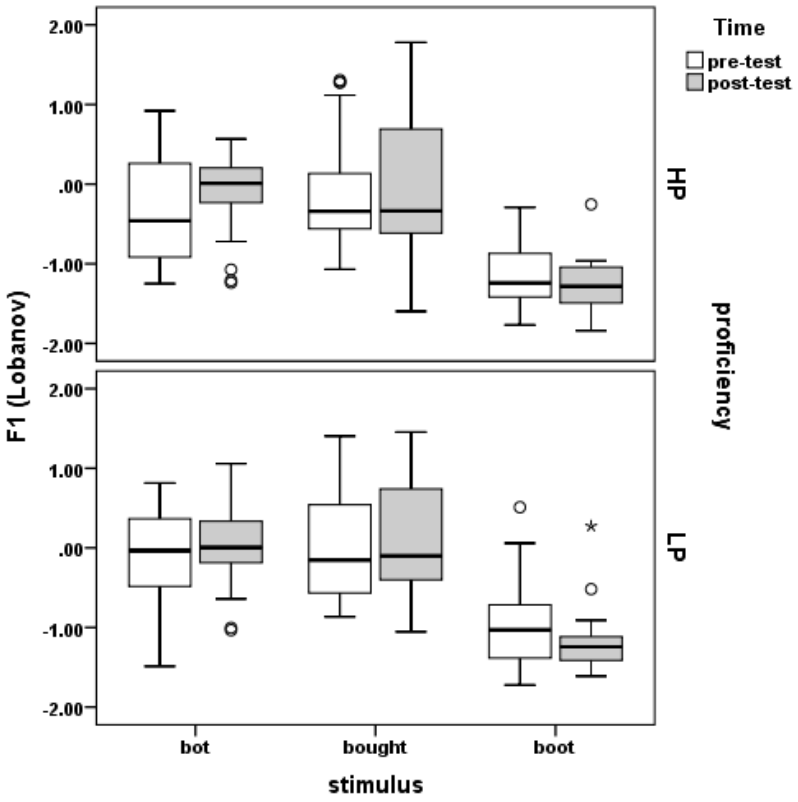


Figure 7.10: Boxplots showing F1 values for vowel group 3 (*bot, bought, boot)* produced by L2 learners [London (N=16; HP=10, LP=6) and SA (N=9; HP=1, LP=8] at the pre- and post-tests. The F1 values for stimuli were the average of the 2 repetitions of each word for each speaker.

The best fitting-model for F1 included proficiency (HP, LP) as a fixed factor and stimulus and participants as random factors with random intercepts. The best-fitting model excluded all other factors and interactions which means that they were not significant for the analysis. The results from the model showed a significant effect of proficiency, χ^2 (1) = 4.4301, p<.05. The planned contrasts indicated a significant difference in F1 values for the HP participants compared to those of the LP participants, b= -0.0717, SE= 0.034, pMCMC<.05. HP participants tended to produce

bot and *bought* with lower F1 values the LP participants, though these effects were small (see Figure 7.10).

For F2 values, the best fitting-model included time (pre-post) and proficiency (HP, LP) as fixed factors and stimulus and participants as random factors with random slopes. There was no significant effect of the factors showing no significant change in F2 values.

7.4.2.1.2 Duration

Group 1: Beat, Bit, Bet, Bert. Figure 7.11 shows the duration of the vowels (*beat, bet, bert, bit*) in the pre- and post-tests produced by L2 learners in the two training environments. As is shown in Figure 7.11, there were some changes in vowel duration from pre- to post-test especially in the vowels produced by the group in SA. In order to investigate any significant change after training, a linear mixed-effects model was built for the duration data. The best fitting model to the data included; training environment (London & SA), time (pre and post), and proficiency (HP, LP) as fixed factors, and participant and stimulus as a random factor with random intercepts.

The main effect of time was significant, $\chi^2(1) = 32.29$, $p<.001$, suggesting a change in the vowel duration from pre- to post-test. The planned contrasts showed a significant change in vowel duration (longer duration) from pre- to post-test, b= 41.69, SE= 10.78, pMCMC<.001, such that after training, speakers used longer values (see Fig 7.11). The main effect of training environment was not significant, $p>.05$. However, there was a significant two-way interaction between time and training environment, $\chi^2(1) = 7.425$, $p<.05$, which suggests that one of the groups' performance changed more from pre- to post-test than the other. Although all participants made a clear distinction between tense and lax vowels at the pre-test, participants produced beat, bet and bert with a longer vowel duration after training

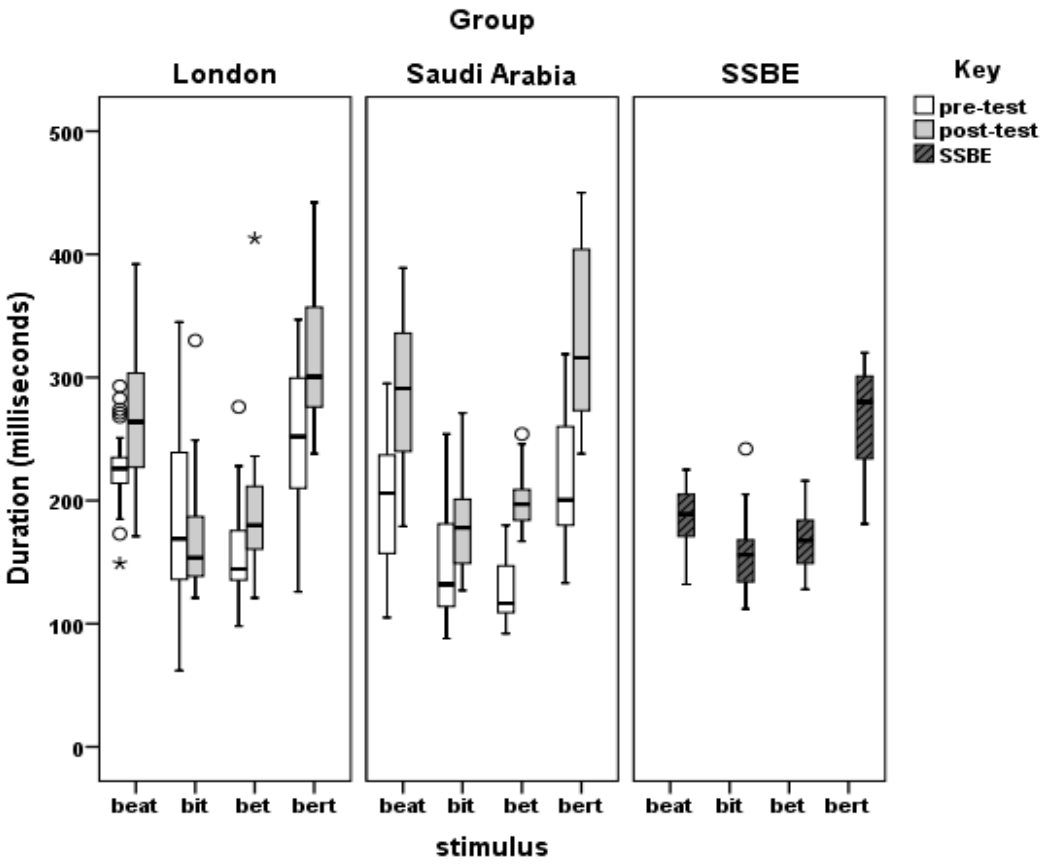


Figure 7.11: Boxplots showing the duration in milliseconds for vowel group 1 (*beat, bit, bet, and bert*) produced by L2 learners in the two training environments (London, N=16, SA, N=9). The duration for stimuli was the average of 2 repetitions of a word for each speaker.

such that it was longer than that of native speakers (see Fig 7.11). The planned contrasts showed that the group that was trained in SA, produced longer vowels from pre- to post-test compared with the equivalent group that was trained in London, b= -28.303, SE= 10.78, pMCMC<.05.

Group 2: Bat, But, Bart. Figure 7.12 displays the vowel duration for (*bat, but, bart*), for participants tested in London and SA. As displayed in Fig 7.12, all participants distinguished between tense and lax vowels at the pre-test, however, the group that was trained in SA appear to change their vowel duration after training, such they made all vowels longer, whilst those in London appeared to make few changes. In order to verify the effect of training environment on vowel duration, a linear mixed-

effects model was built for the duration data based on the duration of the vowels (bat, but, bart) in milliseconds (continuous scale). The best fitting-model was chosen with a top-down approach, and included training environment (London & SA), time (pre-post), and proficiency (HP & LP) as fixed factors, and participant and stimulus as random factors, with random intercepts.

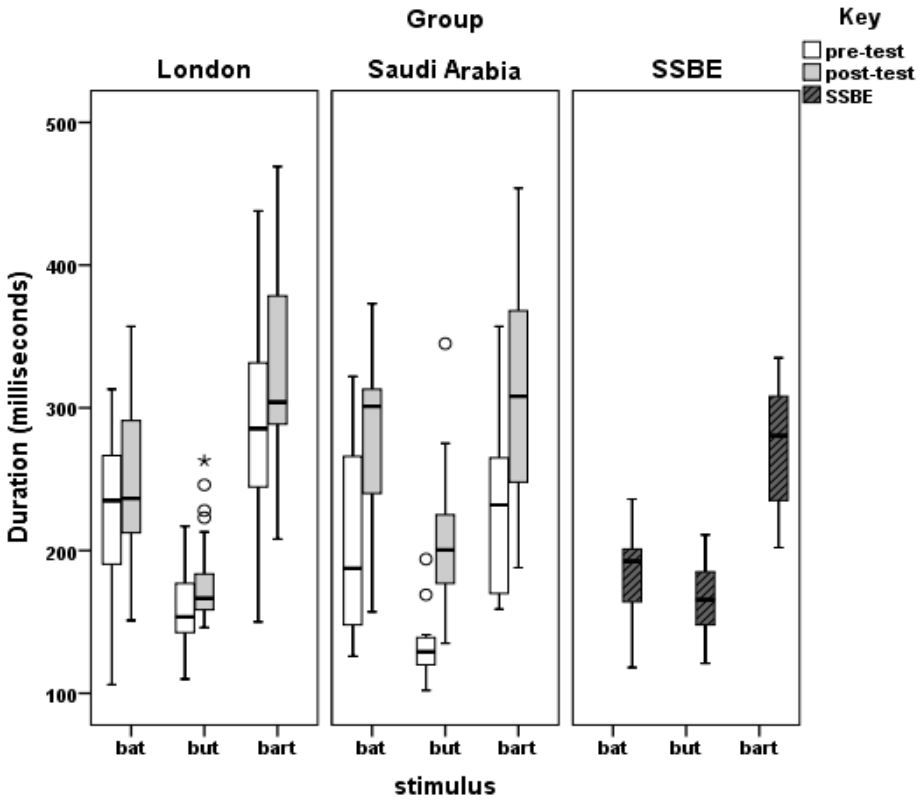


Figure 7.12: Boxplots showing the duration in milliseconds for vowel group 2 (*bat, but, bart*) produced by L2 learners in the two training environments (London, N=16, SA, N=9).The duration for stimuli was the average of 2 repetitions of a word for each speaker.

The best fitting-model excluded the interactions between time, training environment and proficiency, and the interaction between training environment and proficiency, which means that these interactions were not significant for the analysis. The results from the model demonstrated that the main effect of time was significant, χ^2 (1) = 16.552, p<.001, suggesting a change in the vowel duration from pre- to post-test. The planned contrasts showed a significant change in vowel duration from pre- to post-test, b= -26.351, SE= 5.553, pMCMC<.001. The main effect of training

environment was not significant. However, there was a significant two way-interaction between time and training environment χ^2 (1) = 7.120, p<.05, suggesting a possible change from pre- to post-test in one of the group more than the other. The planned contrasts indicated a significant vowel duration change produced by the group that was trained in SA in the post test more than the group that was trained in London, b= 16.38, SE= 6.139, pMCMC<.05. There was no significant effect of proficiency, indicating that the proficiency level did not affect vowel duration change.

Group 3: Bot, Bought, Boot. Figure 7.13 displays the vowel duration at the pre- and post-tests for vowel group 3 (*bot, bought, boot*) for participants tested in London and SA. As displayed in Fig 7.17, all participants could distinguish between tense and lax vowels at the pre-test, however, they appear to change vowel duration for these vowels after training, with all participants producing longer vowels after training.

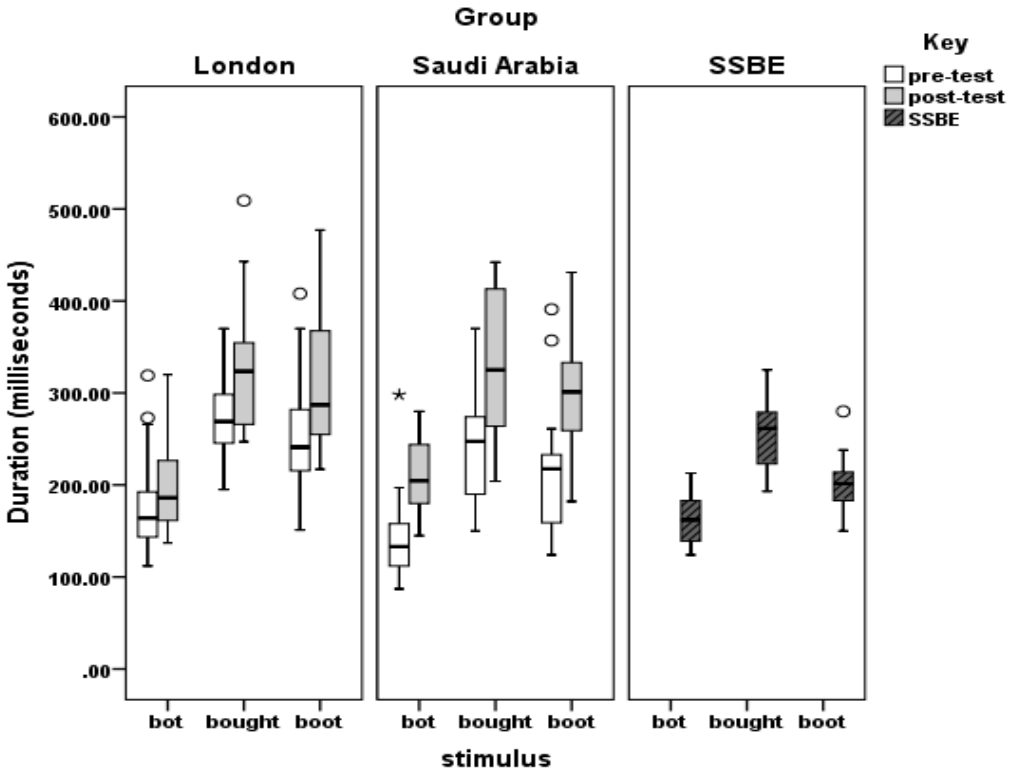


Figure 7.13: Boxplots showing the duration in milliseconds for vowel group 3 (*bot, bought, boot*) produced by L2 learners in the two training environments (London, N=16, SA, N=9). The duration for stimuli was the average of 2 repetitions of a word for each speaker.

To verify any changes in vowel duration after training, a linear mixed-effects model was built for the duration data based on the duration of the vowels (*bot, bought, boot*) in milliseconds (continuous scale). The best-fitting model was chosen with top-down approach and included training environment (London & SA), time (pre-post), and the interaction between time and training environment as fixed factors, and participant and stimulus as random factors, with random intercepts. The best-fitting model excluded all other insignificant factors for the analysis; proficiency, the interaction between training environment and proficiency, and time and proficiency.

The main effect of time was significant, χ^2 (1) = 23.565, p<.001, which suggests a change in vowel duration values from pre- to post test. The planned contrasts showed a significant change (i.e., longer vowel duration) from pre- to post-test, b= -30.56, SE= 6.015, pMCMC<.001, such that all learners used longer vowel duration after training (see Fig 7.13). The main effect of the training environment was not significant, and there was no significant effect of the interaction between time and training environment p>.05.

Summary. Both groups in different training environments (London and SA) changed their vowel production after training. Regarding the spectral changes, though this was not statistically significant, both groups appeared to make some changes to their F1 values for the vowel contrast /e/-/ɪ/ and produced /ɜː/ as a more central vowel, to better match native speakers. Participants in SA also made some subtle changes to F2 but not F1 values for *but* and *bart*, and slight changes in F1 values for the LP learners in both groups for *boot* and *bought.*

Although participants in both training environment groups could distinguish between tense and lax vowels at the pre-test, they made some changes to vowel duration after training. Namely, all participants tended to lengthen all vowels such that they maintained the distinction between tense and lax vowels, but produced these vowels with a longer duration than native speakers. In particular, both groups produced *bet* and *bert* with longer duration at the post-test. However, participants who were trained in SA produced *bart, but, bought* and *boot* with longer duration at the post-test than those who were trained in London.

7.4.2.2 Vowel intelligibility and Goodness Ratings

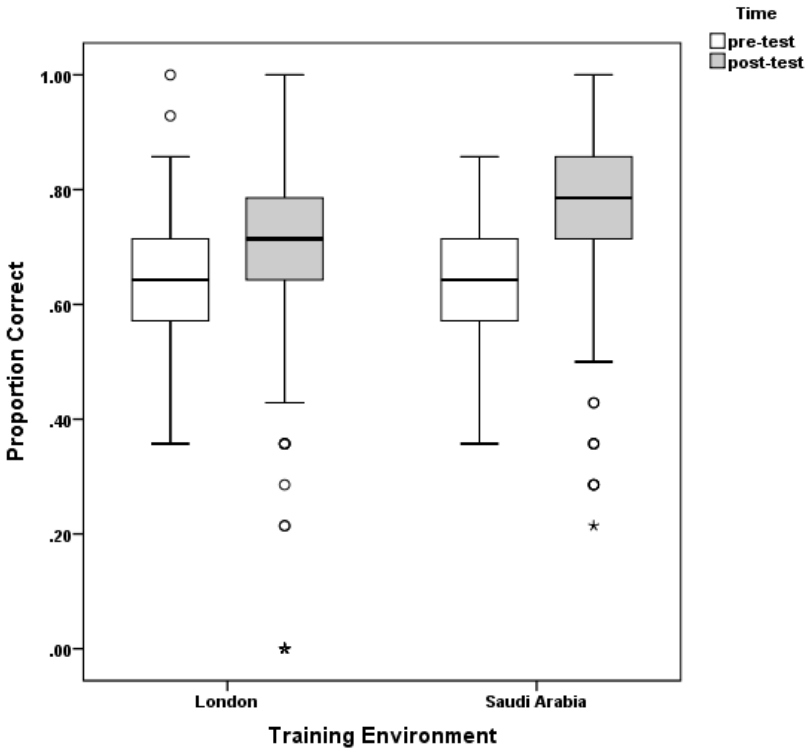


Figure 7.14: Boxplots showing the proportion correct identification for vowels produced by L2 speakers, split by training environment (London, N=16 and SA, N=9).

/b/-V/t/ recordings. As shown in Figure 7.14 all L2 speakers appeared to be more intelligible after training, regardless of the training environment. To verify any significant changes after training, a logistic mixed-effects model was built for the identification data based on binomial responses (correct/incorrect). The best fitting model included time (pre- post), training environment (London, SA), and proficiency (HP, LP) as fixed factors, and participant and stimulus as random factors with a random slope of time with the speaker nested in to the stimulus. This was done because different stimuli were produced by different speakers. The best fitting-model excluded the interactions between time and group, time and proficiency, group and proficiency, and the three-way interaction between time, group and proficiency, which means that these interactions are not significant for the analysis.

There was a significant main effect of time, χ^2 (1) = 8.615, p < .05. This indicates that there was a change in intelligibility from pre- to post-test. The planned contrasts for pre- and post-test identification scores showed a significant improvement at the post test which means that all participants were more intelligible after training, b= -0.355, SE= 0.1209, z= -2.935, p<.05.

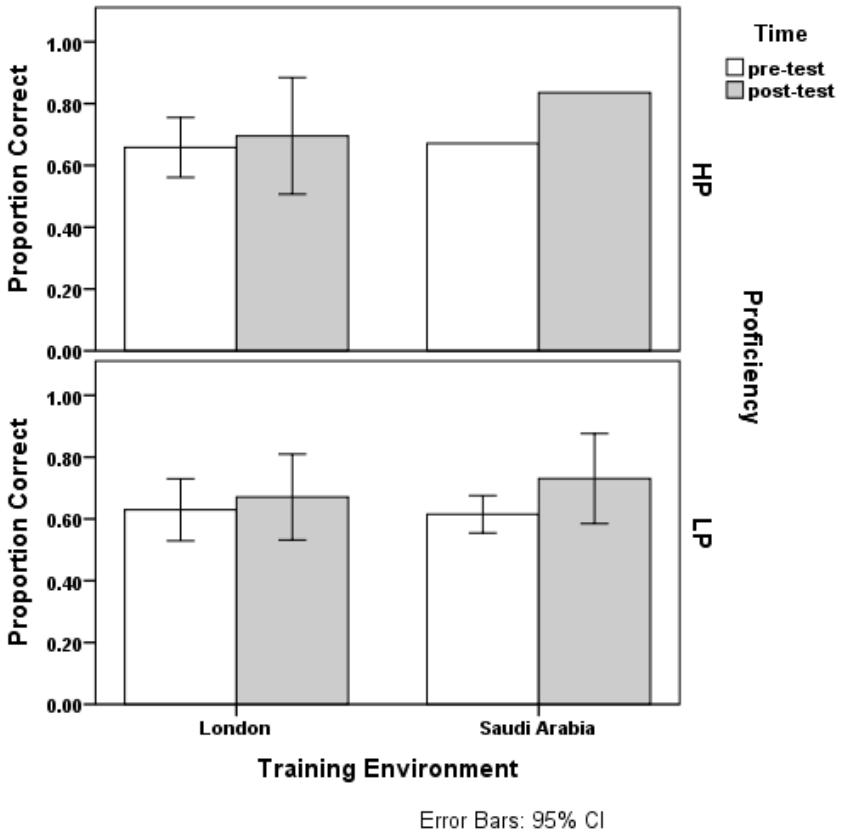


Figure 7.15: Bar chart showing the proportion correct identification for vowels produced by L2; in the two training environments, London & Saudi Arabia, and split by Proficiency Level; High iency (HP; London =10, SA = 1) and Low Proficiency, (LP; London=6 SA=8).

There was also a significant effect of proficiency χ^2 (1) = 4.035, p < .05, suggesting that speakers with different proficiency levels differed in their intelligibility. The planned contrasts showed that overall, the HP participants were more intelligible than the LP participants, b= 0.2208, SE= 0.1099, z= 2.009, p<.05 (see Fig 7.15).

In order to investigate whether the improvement in overall intelligibility was linked to changes in intelligibility for any particular vowel, confusion matrices for pre- and post-tests were calculated (see Appendices 4, 5, 6, and 7). The confusion matrices showed that learners' productions of *bit*, *bet* and *bought* in particular, were better identified after training. For the group that was trained in London, intelligibility for these vowels improved on average by 19% for *bit*, 21% for *bet*, and 26% for *bought*. Intelligibility for the group that was trained in Saudi Arabia improved on average by 20% for *bit*, 25% for *bet*, and 9% for *bought* (see Tables 7.1 & 7.2 for amount of improvement).

		response													
		bait	**bart**	**bat**	**beat**	**bert**	**bet**	**bit**	**bite**	**boat**	**boot**	**bot**	**bought**	**bout**	**but**
	bait	39	0	-2	0	-6	-11	0	-20	0	0	0	0	0	0
	bart	-1	-8	3	0	-2	0	0	0	0	1	-3	6	2	2
	bat	1	1	9	0	-7	0	1	0	0	0	0	-1	0	-4
	beat	2	0	-1	4	0	1	-4	-2	0	0	0	0	0	0
	bert	-1	-1	3	-9	-7	12	0	0	0	0	0	0	1	1
	bet	0	0	11	-18	1	26	-19	-1	0	0	0	0	0	0
	bit	0	0	-2	13	0	-9	21	-22	0	0	0	0	0	-1
	bite	-1	0	0	0	0	-7	-3	10	0	0	0	0	0	1
	boat	0	0	0	0	0	0	0	0	16	-11	-8	-4	4	3
	boot	0	0	0	0	0	0	0	0	-9	24	-1	-12	-1	-1
	bot	0	0	1	0	0	0	0	0	-2	-14	6	7	-1	4
	bought	0	-1	0	0	0	0	0	0	-11	1	-9	9	10	1
	bout	-1	6	0	0	0	0	0	0	-12	-11	-9	4	23	0
stimulus	**but**	0	3	-3	0	1	0	0	0	0	0	1	0	1	-3

Table 7.1: Confusion matrix showing the amount of improvement in percentage correct of the vowel intelligibility for L2 learners who were tested in Saudi Arabia.

		response													
		bait	**bart**	**bat**	**beat**	**bert**	**bet**	**bit**	**bite**	**boat**	**boot**	**bot**	**bought**	**bout**	**but**
	bait	-4	-1	0	0	0	1	-1	4	0	0	0	0	0	0
	bart	-7	-3	1	0	-1	0	-1	0	1	0	4	13	-2	-5
	bat	-1	1	4	1	-3	3	0	0	0	0	-1	0	0	-4
	beat	-3	0	1	-1	1	6	-4	0	-1	1	0	0	0	0
	bert	1	6	1	-3	-12	1	1	0	6	0	0	0	-1	0
	bet	-2	-1	5	6	3	21	-36	0	1	0	0	0	0	3
	bit	0	0	-3	5	0	-14	19	-8	0	0	0	0	0	0
	bite	0	1	0	-1	0	-3	-3	4	0	0	0	0	1	0
	boat	0	0	0	0	-1	0	0	0	-2	3	3	0	0	-3
	boot	0	0	0	0	0	0	0	0	-13	18	-4	1	1	-2
	bot	0	0	-1	0	0	1	0	0	-7	4	4	11	0	-12
	bought	0	-1	-1	0	1	0	0	0	-37	0	-1	26	8	5
	bout	0	0	-1	0	-1	0	0	0	-14	0	-4	-3	17	6
stimulus	**but**	0	1	6	0	0	0	0	0	-1	0	-1	-1	-1	-3

Table 7.2: Confusion matrix showing the amount of improvement in percentage correct of the vowel intelligibility for L2 learners who were tested in London.

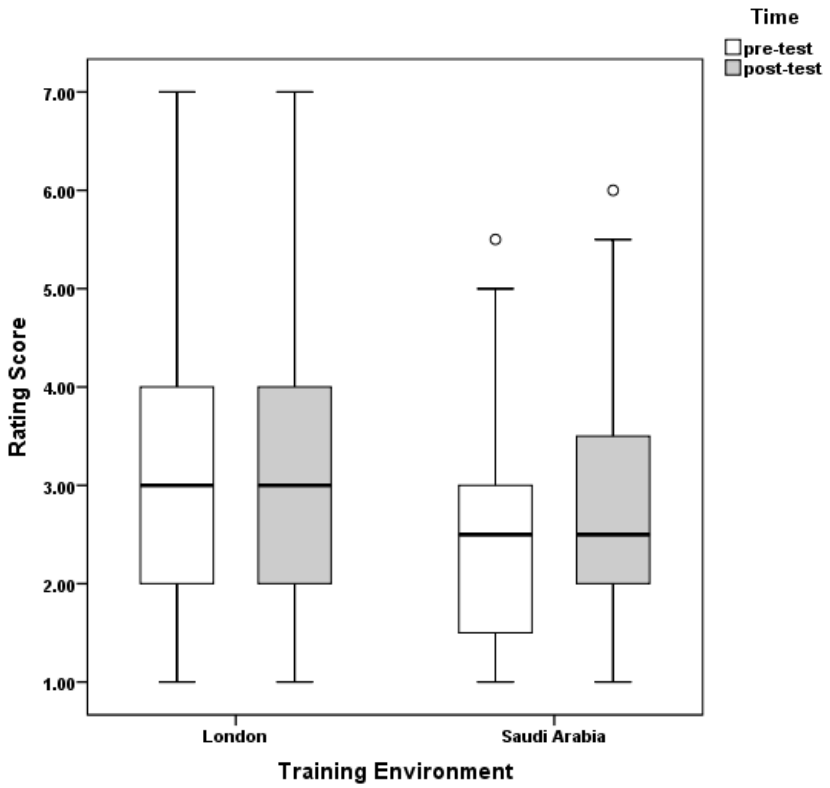


Figure 7.16: Boxplots showing the rating scores for L2 speakers from the production training in the two environments (London, N=16, and in SA, N=9), and rated by SSBE listeners.

Goodness Ratings. In order to investigate whether the ratings were reliable, a reliability test was run on the scores that were given by the 10 raters to the snippets from the pre- and post-test using the using Intra-class Correlation Coefficient (ICC). A two-way mixed model was chosen with "Absolute Agreement" type, and with raters as fixed components to test the level of raters agreement (i.e., whether the raters used the scale in the same or similar way). The results demonstrated a strong consistency in the ratings amongst the raters, Cronbach's Alpha α=.876 which indicated a strong consistency/agreement in ratings amongst the raters (given the fact that a perfect Cronbach's Alpha=1). Average rating scores for each speaker were then calculated and these scores were used in all future analyses.

A linear mixed-effects model was built for the average rating scores. The best-fitting model included proficiency (HP, LP) as a fixed factor, and participant (rater)

and speaker as random factors with random intercept. The best fitting model excluded all other factors (i.e., time, training environment, and the interactions between time and training environment and training environment and proficiency), indicating these interactions were not significant for the analysis. The results from the model indicated no significant main effects for any of the factors, which suggests that there was no significant difference between the accent ratings between the two groups (London & SA), and that these did not change from the pre- to post-test (see Fig. 7.16).

7.4.3 The relationship between vowel identification and vowel intelligibility

In order to investigate any possible relationship between learners' perception (i.e., their scores in the vowel identification task), and their vowel intelligibility (i.e., how accurately English native speakers identify vowels produced by L2 learners), separate correlations were conducted for each training group (London & SA). Results for the London group were presented in Chapter 5, but are summarized here for ease of reference.

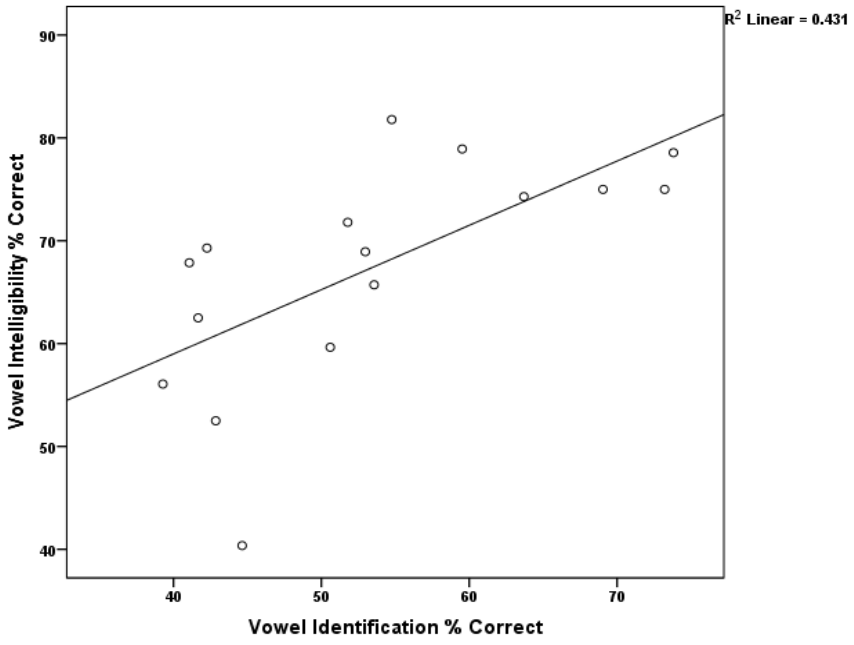


Figure 7.17: Scatterplot of the correlation between vowel identification in percent correct (averaged across pre & post-tests), and the vowel intelligibility in percent correct identified by SSBE listeners (N=10) for L2 learners' vowels in production group in London (N=16)

As demonstrated in Chapter 5, there was a positive relationship between the vowel identification and the vowel intelligibility for the production group in London. A Pearson correlation indicated that there was a positive relationship between vowel identification and intelligibility, [r=.657, p<.05, R2=.431], indicating that learners who performed well at the vowel identification were more intelligible and vice versa (see Fig. 7.17).

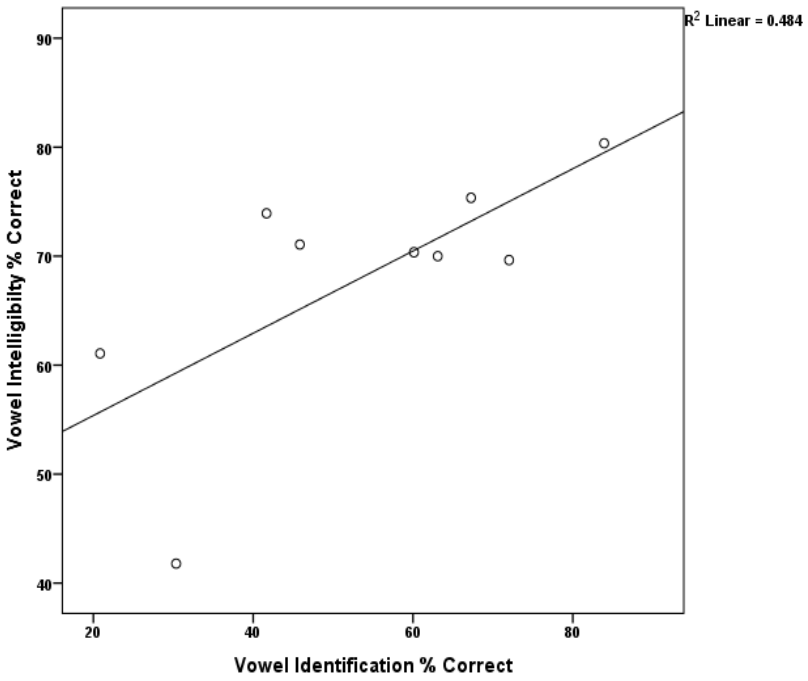


Figure 7.18: Scatterplot of the correlation between vowel identification in percent correct (averaged across pre-post-tests), and the vowel intelligibility in percent correct identified by SSBE listeners (N=10) for L2 learners' vowels in production group in Saudi Arabia (N=9).

For the production group in Saudi Arabia, a Pearson correlation also indicated a significant correlation between the vowel identification and the vowel intelligibility, [r=.696, p<.05, R2= 0.484], (see Fig 7.18). These correlations confirm that perception and production are somehow related; those learners who performed better in vowel perception were more intelligible to native speakers, and those who performed worse were less intelligible.

7.5 Discussion

The present study investigated whether training environment (immersion vs. non-immersion) affects the efficiency of production training. The results from the experiments revealed that both groups improved after training to some extent, though the group that was trained in a non-immersion setting (i.e., in Saudi Arabia) improved in the vowel identification and speech recognition in noise tasks, whilst the group that was trained in London did not. Interestingly, this contradicts the conclusion drawn in Chapter 5 that phonetic training seems to be domain specific (i.e., production training improves production but not perception and perception training improves perception but not production). Instead, it appears that learning environment has a role in what is learned from training.

This supports the notion that not only natural exposure to speech improves performance in perception, but that there is also some aspect of directing or focusing learner's attention to phonetic differences in the production training that is beneficial for speech perception. Perhaps it is the case that because the SA group did not have regular interactions with English speakers, they used the production training as a more holistic tool for acquiring English than did the London group. That is, perhaps the SA group's attention was directed towards using this training as a tool to improving their production and perception whereas the London group just used it as a way to adjust their motor patterns to improve production.

Another possibility is that because participants in SA were mostly recruited from a language institute, they may have been keener to learn and improve their English perception and production. In contrast, participants in London were mostly recruited from Brunel University in London, were not studying English and instead, spent a lot of time working independently in laboratory-based research. These participants also reported that they spent a lot of time with other Saudi or Arabic-speaking students and did not interact that much with English speakers. It is possible that at least in terms of improving their production and perception for spoken English, this group of participants were not as motivated as those participants in Saudi Arabia who were studying English.

Another possibility is that participants in SA may have perceived the vowels produced in the training programme as the default way of producing and perceiving English phonemes, and have thought that they would not be intelligible enough if they did not produce and perceive phonemes in this way. Participants from London on the other hand, are more familiar with the experience that even if their speech is not native-like, one can still be understood. Consequently, these London-based participants may not have valued sounding native-like as an important part of learning English, and thus may not have learned as much overall from the training programme.

That being said, both groups appeared to improve their vowel production for certain vowel contrasts. Although not statistically significant, participants changed their production of /ɪ/-/e/; participants in both groups produced /e/ vowel with lower F1 values than /ɪ/, and produced /ɪ/ with higher F1 values than /e/ before training, and altered the categories such that they produced /ɪ/ with a lower F1 and /e/ with a higher F1 after training. That is before training, all participants produced *bit* such that it sounded closer to *bet*, and *bet* such that it sounded like *bit*, but after the training this was reversed such that *bit* was closer to SSBE *bit* and *bet* was closer to SSBE *bet*

On the other hand some vowels, though they differed from SSBE, did not change significantly after training (e.g., /ɒ/, /ɔ/, and /u/), perhaps because these two vowels are very close to participants' L1 vowel /u/. One might expect that they would assimilate these vowels to their native vowel /u/, and perceive them using this vowel category as they did for the English vowel /u/. Instead participants could detect the difference between their L1 category (i.e., /u/) and the L2 categories especially for /ɒ/ and /ɔ/. However, they did not produce /ɒ/ and /ɔ/ using either L1 or L2 category, and they established a new category that does not belong to either an L1 or L2 phonemic categories. This supports the Speech Learning Model theory (SLM; Flege, 1995), that posits that when L2 learners fail to assimilate an L2 category to an L1 category, and are thus unable to produce it like the L2 category, they establish a new category between their L1 and L2 categories.

Lastly, it is possible that the amount of production training was insufficient for large-scale change in production. Participants received 5 sessions of articulatory training and though every effort was made to train all vowels equally, it is possible that learners focussed on contrasts that they found particularly difficult and which they

judged to be important for their intelligibility. It is possible that with more training sessions, learners may have been able to make changes to more vowels. As was argued in Chapter 5, the idea of recording participants in each session might enable measurement of the amount of improvement. This would measure after each session whether their learning increased with the number of the sessions, or whether it reached a ceiling effect of learning by the fifth session. If so, then increasing the number of sessions besides recording all trials in the training sessions could potentially tell us more about the reasons behind the improvements of some but not all vowels.

However, although there were improvements in vowel identification performance, participants did not reliably improve in their category discrimination. They found some vowels very easy to discriminate, and therefore, as performance was high, it is possible that there was not room for improvement. That said, they did find some vowels harder to discriminate than others. One explanation is that participants are better at distinguishing certain categories based on their existing representations, and perform well with these in identification tasks as a result of training, but do not change their underlying representations, i.e., no change in performance in the category discrimination task. This provides additional evidence for the hypothesis that training does not lead to low-level changes in category representations but instead, enables learners to better match their existing representations with those in the L2 (see Iverson & Evans, 2009).

So what is being learnt from production training? The current study provides further evidence for the relationship between speech perception and production. It was argued in Chapter 5 that phonetic training seems to be domain specific (i.e. production training improves production but not perception and perception training improves perception but not necessarily production). This conclusion was based on results from learners who were trained in London on production-based, perception-based and a hybrid programme that combined both perception and production training.

However, the findings from testing L2 learners in Saudi Arabia a non-immersion setting, do not support this conclusion. Learners who were trained on production in Saudi Arabia improved their vowel identification and their speech recognition in noise, as well as vowel production. This suggests that production training yields improvement in perceptual abilities as well as improving production,

and that this may be dependent either on learning environment itself, or perhaps, more likely, learners' motivation for learning. Indeed, though it was not statistically significant, there was a tendency for participants who were trained in SA to change their production to sound more intelligible than participants who were trained in London after training, and these SA participants were likely more motivated to learn than those tested in London. In summary, production training was shown to be beneficial for L2 learners in a non-immersion setting, and depending on their willingness and motivation to learn, production training appears to lead to improvements in perception as well as production.

Chapter 8 General discussion and conclusion

This thesis examined the acquisition of English phonemes by native Arabic speakers with two main goals; 1) to explore difficult phonemes for Arabic learners of English, and 2) to investigate the relationship between their perception and production of those phonemes. Two main studies were conducted, the first to investigate the problematic phonemic contrasts for Arabic learners of English, and the second to investigate the relationship between speech perception and production in relation to phonetic training type. A group of Arabic learners completed either; PT (5 one-to-one articulatory training sessions), HVPT (5 sessions) or a HTP (5 sessions; including one session of the production training, and four sessions of HVPT) training programme, and the effect of different training types on the indirect speech domain (i.e., if individuals were trained on their production will their perception improve as well as their production and vice versa) was investigated. Two other follow-up studies were conducted; one to investigate the retention of learning in all three training types, and to investigate if production and hybrid training, like HVPT, yield long-term learning. The second follow-up study investigated the benefits of production training in different immersion settings (immersion vs. non-immersion) by comparing the perceptual and production changes before and after production training in two groups; one trained in London (immersion setting, the same group as in Study 2), and the other trained in Saudi Arabia (non-immersion setting).

8.1 What kind of phonemes did Arabic speakers find confusable?

Current theoretical accounts offer several explanations for why L2 learners find some L2 phonemes hard to perceive and produce. Most of these studies attribute such difficulties to the relationship between the individual's L1 and the L2 phoneme inventories (e.g., PAM Best et al, 1995; Best and Taylor, 2007; SLM Flege et al., 1995; Iverson et al, 2003), and/or the size of the phoneme inventory of L1 compared to that of L2 (e.g. Iverson and Evans, 2007). This difficulty in learning L2 phonemes is thought to be language-specific. That is, speakers with different L1 backgrounds have different difficult phonemic contrasts (e.g., for Japanese speakers the difficult phoneme contrast is /r/-/l/, and for the Spanish speakers is the contrasts /i/-/ɪ/). However, there was no study to my knowledge when I started designing the experiments that investigated the perception and production of British English

phonemes by adult Arabic learners (though see Shafiro et al., 2012 for American English perception by Arabic speakers). This thesis aimed initially at exploring these difficulties, by testing Arabic speakers on several perceptual and production tasks.

The results in the Study 1 (Chapter 3) demonstrated that Arabic speakers find English vowels more difficult than consonants, though there were some confusions between some consonant contrasts; /ʒ/-/dʒ/, /ʃ/-/tʃ/, /m/-/n/-/ŋ/. The confusions that Arabic speakers make can be explained with regard to the relationship between the phonemic inventory in their L1 compared to that of the L2. That is, they find the phonemes that do not occur in their L1 harder to perceive and produce than those that do occur in their L1. I hypothesised that, given the Arabic consonant numbers (28), and the number of vowels (6), they would find vowels more challenging than the consonants. The results from Study 1 supported my hypothesis; the Arabic participants found vowels more confusable than consonants. This might be explained by the size of the L1 and L2 phoneme inventories; Arabic learners have 28 consonants onto which they can map the English consonants, whereas they have only 6 vowels against the 17 of British English. This may explain why Arabic learners appear to assimilate the vowels that occupy a place in the vowel space that is near their L1 to their nearest L1 category. For example, they assimilate almost all back vowels to the L1 category /u/. Indeed, having a smaller phonemic inventory might not facilitate learning of L2 phonemes as much as the larger phonemic inventory (Iverson and Evans, 2007). Iverson and Evans (2007) trained Spanish speakers who have only 5 vowels in their vowel inventory, and German speakers, who have 18 vowels (15 monophthongs, and 3 diphthongs), with British English vowels using a HVPT training programme. They found that though all learners improved to some extent, German speakers benefitted from training more than did Spanish learners. They argued that the larger German vowel inventory may have facilitated learning of L2 vowels as Germans were able to utilize their native categories which were a better match than those of Spanish to the British English vowel inventory.

Study 1 also provided evidence that Arabic speakers did not rely totally on duration when identifying the vowels with equated duration in noise. This is possibly because there were tested in noise condition, in which they performed poorly even with natural vowels in noise. This may be because their knowledge of L2 cues (i.e.,

F1 &F2) is not robust enough to help them identify vowels in noise. Another possibility is that they were affected by noise and find it hard to identify the vowels in noise, and thus, equating the duration in noise, did not make a big difference to performance since they found the natural vowels in noise hard enough to identify.

Replicating previous studies (e.g., Flege, 1993; Bradlow et al., 1997; Flege and Schmidt, 1995) the results from study 1 also provide evidence for a link between L2 speech perception (i.e., how accurately English speakers identified English vowels) and production (i.e., how accurately SSBE speakers identified vowel productions of Arabic speakers). Due to time restrictions and the fact that Arabic learners have more confusions with vowels, than consonants, only intelligibility for vowels was tested. However, given such a strong correlation between vowel perception and production, it seems reasonable to assume that consonant perception might also be related to L2 production as well.

The link between speech perception and production has been a longstanding debate in L2 literature. As mentioned in Chapter 4 (p., 85) a number of theories propose that there is a strong link between speech perception and production (e.g., Motor theory, Liberman et al., 1967, Liberman and Mattingly, 1985; Direct realist theory, Fowler, 1981, 1986; Best, 1995; General auditory approach, Diehl et al., 2004). However, evidence for such a link is not always clear, especially in the phonetic training studies. Previous studies have investigated the effect of HVPT (perception-based training) on perception and production, and some have found that training perception leads to improvements in production (e.g., Bradlow et al, 1997), while others have found little or no relationship between training perception and improving in the production domain. For instance, Hattori (2009) found that training production did not lead to improvement in perception, concluding that perception and production may have independent underlying representations.

The other main goal of this thesis is to further explore the relationship between L2 speech perception and production by investigating the effect of different phonetic training types on Arabic learners of English. Arabic speakers were assigned randomly to one of three training programmes; PT, HVPT, and a HTP programme. The aim was to investigate the effect of training one speech domain on the improvement of the other (i.e., whether training perception improve production and vice versa). The hypothesis

was that the hybrid programme might produce more robust learning in both domains compared to the HVPT and the production training, given that it trains both speech domains.

8.2 What has been actually learned after training?

The results from Study 2 support the hypothesis, and learners who were trained in the hybrid program improved both their perception and production of some vowels (e.g., /ɪ/-/e/ contrast). That being said, training perception seems to improve perception, but not production (production measured acoustically), and production seems to improve production (acoustic measures) but not perception (vowel identification). It therefore seems reasonable to argue that training is largely domain specific.

That being said, participants in both production and hybrid programs, improved in their production of only certain vowel contrasts, namely /ɪ/-/e/, rather than improving all trained vowels. It is possible that this is because the difference between /ɪ/ and /e/ can be visualised by looking in a mirror, i.e., seeing the jaw drop for production of /e/. While there is a similar difference for /ɔ/ and /ɒ/ (i.e., a difference in F1), participants find it more difficult to acquire the subtle change between /ɔ/ and /ɒ/ and the amount of jaw movement is smaller. Another explanation is that they did not receive enough training to improve on all vowels. In my study, individuals were trained on 14 English vowels over 5, 40-minute sessions. In contrast, Hattori (2009) trained Japanese speakers on the production of English /r/ and /l/, using only three minimal pair words (*lack, rack, lick, rick, loom, room*), using real-time spectrograms, and over ten sessions. After training, he found large improvements in production such that they had achieved native-like production of the /r/-/l/ contrast. For such a large number of contrasts, perhaps five sessions of training is not enough for learners to make robust adjustments to their production. Given that for the vowels, it has been shown that perceptual training for vowels is more effective when all vowels rather than a subset are trained (Nishi and Kewley-Port, 2007), increasing the number of training sessions rather than training a smaller number of vowels, may lead to more learning.

Although the results of Study 2 suggest that training is largely domain specific, I find it hard to conclude that speech perception and production have independent underlying category representations. Previous perceptual training studies (e.g., Bradlow et al, 1997), have found that perceptual training can lead to improvements in production as well as improvement perception. In these studies, improvements in production were based on improvements in intelligibility measures (i.e., native speakers identifying L2 speakers' productions). It is possible that such measures may tell us more about potential improvements in production than acoustic measures, such as the F1, F2 and duration measures presented here. This led me to test L2 speakers in study 2 for their intelligibility, to investigate whether the native English speakers would judge the production after training to be more intelligible. The results showed that all participants were more intelligible after training regardless of the training type. This suggests that all training programmes, including HVPT, led to improvements in participants' vowel production.

Individuals who were trained in an immersion setting using production training did not improve in their vowel identification accuracy, in contrast to those in the HVPT and HTP conditions, who improved after training. As mentioned before in Chapter 4, the HVPT training programme itself uses a task which is much like the vowel identification task with feedback. It is possible that repeated exposure to this kind of task in both the HVPT and HTP programmes enabled individuals to become better at mapping their own underlying representations onto the English stimuli, and that this enabled them to improve in the vowel identification task. This was supported by the results from category discrimination task, in which there were only small changes to L2 category discrimination accuracy. This further supports Iverson and Evans (2009), who claim that auditory training improves the ability of individuals to apply their existing category knowledge of both L1 and L2 categories but without changing those category representations (e.g., use of cues).

The proficiency level of the L2 learners was also found to affect learning in some tasks. In vowel identification, LP learners in the HVPT and HT training groups improved more than HP learners, perhaps because they had more room to learn: LP learners started with a lower identification score and one could imagine that it is to improve from a lower than a higher starting point. However, in speech in noise, HP

learners improved more than the LP speakers regardless of the training type. Since proficiency in these studies was determined by a grammar test, this suggests that individuals need a certain level of grammatical and lexical knowledge to apply learning on isolated sounds to a real-world context.

8.3 Long-term learning

In line with previous studies (e.g., Bradlow et al., 1999), HVPT was shown to yield long-term learning for vowel identification; the results from Study 3 demonstrate that the HVPT group retain learning 6 months after training. Yet, their vowel production did not improve. In Study 3, I found that the PT and HTP training programmes also lead to retention of learning in both perception and production and interestingly, that there was evidence for further learning. This wasn't surprising in some ways, as the learners were tested in an immersion setting where they were regularly exposed to their L2, English, but hasn't been shown in previous studies. In particular, learners who had completed the PT programme continued to improve after training and this was affected by proficiency; the LP participants in this group improved more at the retention test than the HP learners. This is possibly because LP participants had more scope for learning, but interestingly, it might also suggest that production training served a key role in perceptual learning; production training may have enabled learners to redirect their attention to the difference between certain categories.

Participants appear not only to retain their performance in speech recognition in noise, but also performed better at the retention test, especially the LP participants. As mentioned above, since the proficiency is based on a grammar test and this real-world task needs lexical and grammatical knowledge, after 6 months, individuals might learn through more exposure to L2, or through their studies (all participants were university students in London), gaining more lexical and grammatical knowledge that they can apply in this task.

8.4 The effect of immersion settings on learning

The hypothesis behind Study 4 was that L2 learners in an immersion setting might improve more than those who were trained in a non-immersion setting, arguably because the L2 learners in the immersion settings are indirectly trained through daily

interactions with native English speakers. However, the results from Study 4 did not support this hypothesis; participants in a non-immersion setting improved both their perception and production after PT training whereas those in the immersion setting improved only in production. Furthermore, SSBE listeners found the participants who were trained in a non-immersion settings more intelligible than those who were trained in London at the post-test. This improvement was further supported by the correlation between the native listeners' identification of the L2 speakers' vowel production and the vowel identification. These results contradict previous studies. For example, Iverson et al (2012), trained French speakers in London and French speakers in France using HVPT and found similar improvements for both groups in perception and production improvement. Although in Study 4 the training type was different from that used in Iverson et al. (2011), a similar conclusion may apply here; that it is not the exposure per se to natural speech that improves performance in speech production, but just exposure itself. Indeed, there appears to be some benefit of directing the learners' attention to certain phonetic differences that helps them improve their L2 vowel perception and production regardless of learning environment. However, somehow, the exposure for the participants in Study 4 that were trained in London was not a bonus for overall learning. Though they improved their production of certain vowels (e.g., /ɪ/-/e/), as a group, they did not show reliable improvements in speech perception. The participants who were trained in a non-immersion settings on the other hand, improved in both speech domains. One possibility is that they were more motivated to learn, since most of them were recruited from a language centre, where they pay privately to learn English for academic or business purposes.

8.5 Summary

Overall, the findings that emerged from this research bring a substantial contribution to our understanding of the problematic contrasts for Arabic learners of English, and shed further light on the nature of the link between speech perception and production with regard to L2 training. Although I found that training seems to be largely domain specific, when L2 learners' production was judged by SSBE listeners, there was a link between accuracy in production and intelligibility. The thesis also developed and tested a hybrid training program, combining training in production and perception. Based on the combined evidence from the training study, Study 2, retention study and Study 4, production training appears to lead to a deeper level of

learning and thus, combining production and perception training would seem to be highly beneficial. Arguably, production training involves perception and so this may be why this training programme was successful with the learners in a non-immersion setting, leading to improvements in both production and perception in this particular group.

Lastly, one of the main contribution of this thesis is the design of CALVin (computer assisted learning for vowels interface), which can be used as a teaching tool for phonetics as well as second language learning, enabling naive learners to make direct links between the articulators and the resulting sound.

8.6 Limitations and future research

One limitation with regard to the evaluation of the PT programme, is the lack of a retention test for participants in the non-immersion setting. This was mainly because of time restrictions and practical considerations, as this would have required another trip to Saudi Arabia to re-test the participants. However, given that production training leads to long-term learning in the group that was trained in London, it seems plausible to predict that participants in Saudi would retain learning after a while of production training.

The rationale behind the HTP program was to investigate what benefits articulatory training can add to HVPT. Given that the HTP programme consisted of one session of PT and four sessions of HVPT, yet participants improved in their production of some of vowels, it would be interesting to investigate whether equalizing the session numbers in the training would lead to improvement of more vowels than did just one session of PT. Indeed, it would be interesting for future work to develop a training programme that includes intensive training in both speech domains. Finally, future work could include more conversation-like tasks to assess the effectiveness of training beyond word-level identification and production.

Bibliography

Abboud, P. F., & McCarus, E. N. (Eds.). (1983). Elementary Modern Standard Arabic: Volume 1, Pronunciation and Writing; Lessons 1-30 (Vol. 1). Cambridge University Press.

Abd-El-Jawad, H. R. (1987). Cross-dialectal variation in Arabic: Competing prestigious forms. Language in Society, 16(03), 359-367.

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.

Akahane-Yamada, R., Kato, H., Adachi, T., Watanabe, H., Komaki, R., Kubo, R.,& Kawahara, H. (2004). ATR CALL: A speech perception/production training system utilizing speech technology. In 18th Internat. Congress on Acoustics, Proc. ICA (Vol. 3, pp. 2319-2320).

Akahane-Yamada, R., Tohkura, Y. I., Bradlow, A. R., & Pisoni, D. B. (1996). Does training in speech perception modify speech production?. In Spoken Language, 1996. ICSLP 96. Proceedings. Fourth International Conference on (Vol. 2, pp. 606-609). IEEE.

Al-Ani, S. (1978). The development and distribution of the Qaaf in Iraq. Readings in Arabic linguistics. Bloomington: Indiana University Linguistics Club, 103-12.

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

Al-Tamimi, J. (2007, August). Static and dynamic cues in vowel production: A cross dialectal study in Jordanian and Moroccan Arabic. In Proc. of the 16th International Congress of Phonetic Sciences (ICPhS), Saarbrücken, Germany.

Amayreh, M. M. (2003). Completion of the consonant inventory of Arabic. Journal of speech, language, and hearing research : JSLHR, 46(3), 517–29.

Amayreh, M. M., & Dyson, A. T. (1998). The acquisition of Arabic consonants. Journal of Speech, Language, and Hearing Research, 41(3), 642-653.

Ananthakrishnan, K. S. (2003). Computer aided pronunciation system (CAPS) (Doctoral dissertation, University of South Australia).

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., & Weinert, R. (1991). The HCRC map task corpus. Language and speech, 34(4), 351-366.

Antoniou, M., Best, C. T., & Tyler, M. D. (2013). Focusing the lens of language experience: Perception of Ma'di stops by Greek and English bilinguals and monolinguals. The Journal of the Acoustical Society of America, 133(4), 2397.

Antoniou, M., Best, C. T., Tyler, M. D., & Kroos, C. (2010). Language context elicits native-like stop voicing in early bilinguals' productions in both L1 and L2. Journal of phonetics, 38(4), 640–653.

Antoniou, M., Best, C. T., Tyler, M. D., & Kroos, C. (2011). Inter-language interference in VOT production by L2-dominant bilinguals: Asymmetries in phonetic code-switching. Journal of phonetics, 39(4), 558–570.

Antoniou, M., Tyler, M. D., & Best, C. T. (2012). Two ways to listen : Do L2-dominant bilinguals perceive stop voicing according to language mode ? Journal of Phonetics, 40(4), 582–594.

Aoyama, K., Flege, J. E., Guion, S. G., Akahane-Yamada, R., & Yamada, T. (2004). Perceived phonetic dissimilarity and L2 speech learning: the case of Japanese /r/ and English /l/ and /r/. Journal of Phonetics, 32(2), 233–250.

Baayen, R. H. (2008). Analyzing linguistic data: A practical introduction to statistics using R.

Badin, P., Tarabalka, Y., Elisei, F., & Bailly, G. (2010). Can you "read" tongue movements? Evaluation of the contribution of tongue display to speech understanding. Speech Communication, 52(6), 493–503.

Bailey, P. J., & Haggard, M. P. (1973). Perception and production: Some correlations on voicing of an initial stop. Language and speech, 16(3), 189-195.

Bailey, P. J., & Haggard, M. P. (1980). Perception-production relations in the voicing contrast for initial stops in 3-year-olds. Phonetica, 37(5-6), 377-396.

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

Baker, W., & Trofimovich, P. (2006). Perceptual paths to accurate production of L2 vowels: The role of individual differences. IRAL – International Review of Applied Linguistics in Language Teaching, 44(3), 231–250.

Bani-Yassin, R. and Owens, J. (1987). "The Phonology of a Northern Jordanian Arabic Dialect". Zeitschrift der Deutschen Morgenlandischen Gesellschaft, 137:2, pp. 297-331.

Barkat-Defradas, M., Al-Tamimi, J. E., & Benkirane, T. (2003). Phonetic variation in production and perception of speech: a comparative study of two Arabic dialects. Solé, Recasens, Romero, Proc. 15th Int. Congr. Phonet. Sci., Barcelona, 857-860.

Barreda, S., & Nearey, T. M. (2013). Training listeners to report the acoustic correlate of formant-frequency scaling using synthetic voices. Dimension Contemporary German Arts And Letters, 133(November 2012), 1065–1077.

Bent, T., & Bradlow, A. R. (2003). The interlanguage speech intelligibility benefit. The Journal of the Acoustical Society of America, 1143), 1600.

Bent, T., & Holt, R. F. (2013). The influence of talker and foreign-accent variability on spoken word identification. Perception, 133(March), 1677–1686.

Best, C. C., & McRoberts, G. W. (2003). Infant perception of non-native consonant contrasts that adults assimilate in different ways. Language and speech, 46(2-3), 183-216.

Best, C. T. (1994). The emergence of native-language phonological influences in infants: A perceptual assimilation model. The development of speech perception: The transition from speech sounds to spoken words, 167, 224.

Best, C. T. (1995). A direct realist view of cross-language speech perception. Speech perception and linguistic experience: Issues in cross-language research.

Best, C. T., & Halle, P. A. (2010). Perception of initial obstruent voicing is influenced by gestural organization. Journal of Phonetics, 38, 109–126. doi:10.1016/j.wocn.2009.09.001

Best, C. T., & McRoberts, G. W. (2003). "Infant perception of non-native consonant contrasts

Best, C. T., & Strange, W. (1992). Effects of Phonological and Phonetic Factors on Cross- Language Perception of Approximants *. English, 89-108.

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. Language experience in second language speech learning: In honour of James Emil Flege, 13-34.

Best, C. T., Bradlow, A. R., Guion-anderson, S., & Polka, L. (2011). Using the lens of phonetic experience to resolve phonological forms. Journal of Phonetics, 39(4), 453–455. doi:10.1016/j.wocn.2011.08.006

Best, C. T., McRoberts, G. W., & Goodell, E. (2001). Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. The Journal of the Acoustical Society of America, 1092), 775.

Best, C. T., McRoberts, G. W., & Sithole, N. M. (1988). Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants. Journal of Experimental Psychology: Human perception and performance, 14(3), 345.

Birdsong, D., Bohn, O. S., & Munro, M. J. (2007). Nativelike pronunciation among late learners of French as a second language. Language experience in second language speech learning, 99-116.

Boersma, P. (2009). Cue constraints and their interactions in phonological perception and production *, (July), 1–42.

Boersma, P., & Weenink, D. (2005). Praat: doing phonetics by computer (version 4.3. 14).[Computer program]. Retrieved May 26, 2005.

Boersma, P., & Weenink, D. (2009). Praat v. 5.1. A system for doing phonetics by computer.

Bohn, O. S. (1995). Cross-language speech perception in adults: First language transfer doesn't tell it all. Speech perception and linguistic experience: Issues in cross-language research, 279-304.

Bohn, O., & Best, C. T. (2011). Native-language phonetic and phonological influences on perception of American English approximants by Danish and German listeners. Journal of Phonetics. doi:10.1016/j.wocn.2011.08.002

Bosch, L., & Sebastián-Gallés, N. (2003). Simultaneous bilingualism and the perception of a language-specific vowel contrast in the first year of life. Language and Speech, 46(2-3), 217-243.

Bosker, H. R., Pinget, a.-F., Quene, H., Sanders, T., & de Jong, N. H. (2012). What makes speech sound fluent? The contributions of pauses, speed and repairs. Language Testing.

Bradlow, a R., Akahane-Yamada, R., Pisoni, D. B., & Tohkura, Y. (1999). Training Japanese listeners to identify English /r/ and /l/: long-term retention of learning in perception and production. Perception & psychophysics, 615), 977-85.

Bradlow, a R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. (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), 2299–310.

Bradlow, A. R. (2008). Training non-native language sound patterns Lessons from training Iapanese adults on the English. Phonology and second language acquisition, 36, 287.

Broersma, M. (2010). Perception of final fricative voicing: native and nonnative listeners' use of vowel duration. The Journal of the Acoustical Society of America, 127(3), 1636-44. doi:10.1121/1.3292996

Brouwer, S., Mitterer, H., & Huettig, F. (2010). Shadowing reduced speech and alignment. The Journal of the Acoustical Society of America, 128(1), EL32-EL37.

Buali, I. (2010). The Perception and Production of/p/in Saudi Gulf Arabic English: A Variationist Perspective (Doctoral dissertation, Concordia University).

Bundgaard-Nielsen, R. L., Best, C. T., & Tyler, M. D. (2010). Vocabulary size matters: The assimilation of second-language Australian English vowels to first-language Japanese vowel categories. Applied Psycholinguistics, 32(01), 51–67.

Burleson, D. F. (2014). Improving Intelligibility of Non-Native Speech with Computer-Assisted Phonological Training. IULC Working Papers, 7.

Burnham, D. (2003). Language specific speech perception and the onset of reading, 573–609. Cambridge, England: Cambridge University Press.

Cebrian, J. (2006). Experience and the use of non-native duration in L2 vowel categorization. Journal of Phonetics, 34(3), 372-387.

Chaney, C. (1988). Acoustic analysis of correct and misarticulated semivowels. Journal of Speech, Language, and Hearing Research, 31(2), 275-287.

Chen, T. H., & Massaro, D. W. (2008). Seeing pitch: visual information for lexical tones of Mandarin-Chinese. The Journal of the Acoustical Society of America, 123(4), 2356–66. doi:10.1121/1.2839004

Chládková, K., & Podlipský, V. J. (2011). Native dialect matters: perceptual assimilation of Dutch vowels by Czech listeners. The Journal of the Acoustical Society of America, 130(4), EL186–92.

Chou, F. (2005). Ya-Ya language box. A portable device for English pronunciation training with speech recognition technologies. Paper presented at proceedings of Interspeech-2005, Lisbon, Portugal (pp. 169–172).

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. a. (2008). Perception of speech reflects optimal use of probabilistic speech cues. Cognition, 108(3), 804–9.

Cooke, M., Lecumberri, M. G., & Barker, J. (2008). The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception. The Journal of the Acoustical Society of America, 123(1), 414-427.

Cowell, M. W. (2005). A Reference Grammar of Syrian Arabic with Audio CD: (based on the Dialect of Damascus). Georgetown University Press.

Cruttenden, A. (2008) Gimson's Pronunciation of English, 7th Edition. Hodder Education.

Cutler, A. (2000). Listening to a second language through the ears of a first. Interpreting, 5(1), 1-23.

Cutler, a., Demuth, K., & McQueen, J. M. (2002). Universality Versus Language-Specificity in Listening to Running Speech. Psychological Science, 13(3), 258–262.

Cutler, A., Weber, A., Smits, R., & Cooper, N. (2004). Patterns of English phoneme confusions by native and non-native listeners. The Journal of the Acoustical Society of America, 116(6), 3668.

Dalby, J., & Kewley-port, D. (1999). Explicit Pronunciation Training Using Automatic Speech Recognition Technology. Calico Journal, 16(3), 425–446.

Davis, M. H., & Gaskell, M. G. (2009). A complementary systems account of word learning: neural and behavioural evidence. Philosophical Transactions of the Royal Society B: Biological Sciences, 364(1536), 3773-3800.

Demenko, G., Wagner, A., Cylwik, N., & Jokisch, O. (2009). An audiovisual feedback system for acquiring L2 pronunciation and L2 prosody. In SLaTE (pp. 113-116). DiCanio, C. T. (2012). Cross-linguistic perception of Itunyoso Trique tone. Journal of Phonetics, 40(5), 672-688.

Diehl, R. L., Lotto, A. J., & Holt, L. L. (2004). Speech perception. Annual review of psychology, 55, 149–79.

Eimas, P. D., Siqueland, E. R., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. Science, 171(3968), 303-306.

El-Sadany, T. A., & Hashish, M. A. (1989). An Arabic morphological system. IBM Systems Journal, 28(4), 600-612.

Engwall, O. (2008). Can audio-visual instructions help learners improve their articulation?-an ultrasound study of short term changes. In INTERSPEECH (pp. 2631-2634).

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.

Engwall, O., & Bälter, O. (2007). Pronunciation feedback from real and virtual language teachers. Computer Assisted Language Learning, 20(3), 235-262.

Engwall, O., Bälter, O., Öster, A. M., & Kjellström, H. (2006). Designing the user interface of the computer-based speech training system ARTUR based on early user tests. Behaviour & Information Technology, 25(4), 353-365.

Escudero, P. (2000). The perception of English vowel contrasts: acoustic cue reliance in the development of new contrasts. In Proceedings of the 4th International Symposium on the Acquisition of Second-Language Speech, New Sounds (pp. 122-131).

Escudero, P. (2001). The role of the input in the development of L1 and L2 sound contrasts: language-specific cue weighting for vowels. In Proceedings of the 25th annual Boston University conference on language development (Vol. 1, pp. 250-261). Somerville, MA: Cascadilla Press.

Escudero, P. (2005). "Linguistic perception and second-language acquisition: Explaining the attainment of optimal phonological categorization," Ph.D. thesis, Utrecht University, The Netherlands.

Escudero, P., & Boersma, P. (2002). The subset problem in L2 perceptual development: Multiple-category assimilation by Dutch learners of Spanish. In Proceedings of the 26th annual Boston University conference on language development (pp. 208-219). Somerville, MA: Cascadilla Press.

Escudero, P., & Boersma, P. (2004). Bridging the gap between L2 speech perception research and phonological theory. Studies in Second Language Acquisition, 26(04), 551-585.

Escudero, P., & Chládková, K. (2010). Spanish listeners' perception of American and Southern British English vowels. The Journal of the Acoustical Society of America, 128(5), EL254-9.

Escudero, P., Benders, T., & Lipski, S. C. (2009). Native, non-native and L2 perceptual cue weighting for Dutch vowels: The case of Dutch, German, and Spanish listeners. Journal of Phonetics, 374), 452-465. Elsevier.

Eskenazi, M. (2009). An overview of spoken language technology for education. Speech Communication, 51(10), 832–844.

Eskenazi, M., & Hansma, S. (1998). The Fluency pronunciation trainer. Proc. Speech Technology in Language Learning, 77-80.

Evans, B. G., & Iverson, P. (2004). Vowel normalization for accent: An investigation of best exemplar locations in northern and southern British English sentences. The Journal of the Acoustical Society of America, 1151), 352.

Fadiga, L., Craighero, L., Buccino, G., & Rizzolatti, G. (2002). Speech listening specifically modulates the excitability of tongue muscles: a TMS study. European Journal of Neuroscience, 15(2), 399-402.

Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G. (2000). Visuomotor neurons: ambiguity of the discharge or "motor" perception? International journal of psychophysiology : official journal of the International Organization of Psychophysiology, 35(2-3), 165–77.

Fagel, S., & Madany, K. (2008). A 3-d virtual head as a tool for speech therapy for children. In Interspeech (pp. 2643-2646).

Fischer, M. H., & Zwaan, R. A. (2008). Embodied language: a review of the role of the motor system in language comprehension. The Quarterly Journal of Experimental Psychology, 61(6), 825-850.

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), 2973–87.

Flege, J. E. (1991). Perception and production: The relevance of phonetic input to L2 phonological learning. Crosscurrents in second language acquisition and linguistic theories, 2, 249-89.

Flege, J. E. (1993). Production and perception of a novel, second-language phonetic contrast. The Journal of the Acoustical Society of America, 93(3), 1589-1608.

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

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

Flege, J. E. (1997). Effects of experience on non-native speakers' production and perception of English vowels Ocke-Schwen Bohn. Journal of Phonetics.

Flege, J. E. (1999). Age of learning and second language speech. Second language acquisition and the critical period hypothesis, 101-131.

Flege, J. E. (2003). Assessing constraints on second-language segmental production and perception. Phonetics and phonology in language comprehension and production: Differences and similarities, 319-355.

Flege, J. E. (2007). Language contact in bilingualism: Phonetic system interactions. Laboratory phonology, 9, 353-382.

Flege, J. E., & Liu, S. (2001). The effect of experience on adults' acquisition of a second language. Studies in second language acquisition, 23(04), 527-552.

Flege, J. E., & MacKay, I. R. (2004). Perceiving vowels in a second language. Studies in second language acquisition, 26(01), 1-34.

Flege, J. E., & Port, R. (1981). Cross-language phonetic interference: Arabic to English. Language and Speech, 24(2), 125-146.

Flege, J. E., & Schmidt, a M. (1995). Native speakers of Spanish show rate-dependent processing of English stop consonants. Phonetica, 52(2), 90–111.

Flege, J. E., Bohn, O. S., & Jang, S. (1997). Effects of experience on non-native speakers' production and perception of English vowels. Journal of Phonetics, 25(4), 437-470.

Flege, J. E., Frieda, E. M., & Nozawa, T. (1997). Amount of native-language (L1) use affects the pronunciation of an L2. Journal of Phonetics, 25(2), 169-186.

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), 2973–87.

Flege, J. E., Munro, M. J., & MacKay, I. R. (1995). Effects of age of second-language learning on the production of English consonants. Speech Communication, 16(1), 1-26.

Flege, J. E., Munro, M. J., & MacKay, I. R. (1995). Factors affecting strength of perceived foreign accent in a second language. The Journal of the Acoustical Society of America, 97(5 Pt 1), 3125-34.

Flege, J. E., Schirru, C., & MacKay, I. R. (2003). Interaction between the native and second language phonetic subsystems. Speech communication, 40(4), 467-491.

Flege, J. E., Yeni-Komshian, G. H., & Liu, S. (1999). Age constraints on second-language acquisition. Journal of memory and language, 41(1), 78-104.

Flege, James Emil, Birdsong, D., Bialystok, E., Mack, M., Sung, H., & Tsukada, K. (2006). Degree of foreign accent in English sentences produced by Korean children and adults. Journal of Phonetics, 34(2), 153–175.

Fowler, C. A. (1981). Production and perception of coarticulation among stressed and unstressed vowels. Journal of Speech, Language, and Hearing Research, 24(1), 127-139.

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. Journal of Phonetics, 14(1), 3-28.

Fowler, C. A., Brown, J. M., Sabadini, L., & Weihing, J. (2003). Memory and Language Rapid access to speech gestures in perception : Evidence from choice and simple response time tasks q. Journal of Memory and Language, 49, 396–413.

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

Galantucci, B., Laboratories, H., Haven, N., & Fowler, C. A. (2006). The motor theory of speech perception reviewed. Psychonomic Bulletin & Review, 13(3), 361–377.

Gass, S. M., Mackey, A., & Pica, T. (1998). The Role of Input and Interaction in Second Language Acquisition Introduction to the Special Issue. The modern language journal, 82(3), 299-307.

Gilichinskaya, Y. D., & Strange, W. (2010). Perceptual assimilation of American English vowels by inexperienced Russian listeners. The Journal of the Acoustical Society of America, 128(2), EL80-5.

Goto, H. (1971). Auditory perception by normal Japanese adults of the sounds “L” and “R”. Neuropsychologia, 9(3), 317-323.

Gottfried, T. L. (1984). Effects of consonant context on the perception of French vowels. Journal of Phonetics.

Goudbeek, M., Cutler, A., & Smits, R. (2008). Supervised and unsupervised learning of multidimensionally varying non-native speech categories. Speech communication, 50(2), 109-125.

Guion, S. G., Flege, J. E., Akahane-Yamada, R., & Pruitt, J. C. (2000). An investigation of current models of second language speech perception: The case of Japanese adults' perception of English consonants. The Journal of the Acoustical Society of America, 107(5), 2711-2724.

Hagiwara, R., Fosnot, S. M., & Alessi, D. M. (2002). Acoustic phonetics in a clinical setting: a case study of /r/-distortion therapy with surgical intervention. Clinical linguistics & phonetics, 16(6), 425–41.

Harnsberger, J. D. (2001). On the relationship between identification and discrimination of non-native nasal consonants. The Journal of the Acoustical Society of America, 110(1), 489-503.

Hattori, K. (2009). Perception and Production of English /r/-/l/ by adult Japanese speakers. Doctoral thesis submitted to 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. The Journal of the Acoustical Society of America, 125(1), 469–79.

Heeren, W. F. L., & Schouten, M. E. H. (2008). Perceptual development of phoneme contrasts: How sensitivity changes along acoustic dimensions that contrast phoneme categories. The Journal of the Acoustical Society of America, 124(4), 2291-2302.

Hillenbrand, J. M., Clark, M. J., & Houde, R. a. (2000). Some effects of duration on vowel recognition. The Journal of the Acoustical Society of America, 108(6), 3013–22.

Hirata, Y. (2004). Training native English speakers to perceive Japanese length contrasts in word versus sentence contexts. The Journal of the Acoustical Society of America, 116(4), 2384-2394.

Hirata, Y., Whitehurst, E., & Cullings, E. (2007). Training native English speakers to identify Japanese vowel length contrast with sentences at varied speaking rates. The Journal of the Acoustical Society of America, 121(6), 3837–45.

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

Holt, L. L., & Lotto, A. J. (2006). Cue weighting in auditory categorization: Implications for first and second language acquisition. The Journal of the Acoustical Society of America, 119(5), 3059.

Hubais, A., & Pillai, S. (2010). An instrumental analysis of English vowels produced by Omanis. Journal of Modern Languages, 20, 1-18.

Huer, M. B. (1989). Acoustic Tracking of Articulation Errors [r]. Journal of Speech and Hearing Disorders, 54(4), 530-534.

International Phonetic Association. (1999). Handbook of the International Phonetic Association (IPA).

Iverson, P., & Evans, B. G. (2007). Auditory training of English vowels for first-language speakers of Spanish and German. English, August), 1625-1628.

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

Iverson, P., Ekanayake, D., Hamann, S., Sennema, A., & Evans, B. G. (2008). Category and perceptual interference in second-language phoneme learning: An examination of English/w/-/v/learning by Sinhala, German, and Dutch speakers. Journal of Experimental Psychology: Human Perception and Performance, 34(5), 1305.

Iverson, P., Hazan, V., & Bannister, K. (2005). Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English/r/-/l/to Japanese adults. The Journal of the Acoustical Society of America, 118(5), 3267-3278.

Iverson, P., Hazan, V., & Bannister, K. (2005). Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English/r/-/l/to Japanese adults. The Journal of the Acoustical Society of America, 118(5), 3267-3278.

Iverson, P., Kuhl, P. K., Akahane-yamada, R., & Diesch, E. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. Cognition, 87, 47–57.

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(01), 145-160.

Jamieson, D. G., & Morosan, D. E. (1989). Training new, nonnative speech contrasts: A comparison of the prototype and perceptual fading techniques. Canadian Journal of Psychology/Revue canadienne de psychologie, 43(1), 88.

Jia, G., & Aaronson, D. (2003). A longitudinal study of Chinese children and adolescents learning English in the United States. Applied Psycholinguistics, 24(01), 131-161.

Jia, G., Strange, W., Wu, Y., Collado, J., & Guan, Q. (2006). Perception and production of English vowels by Mandarin speakers: Age-related differences vary with amount of L2 exposure. The Journal of the Acoustical Society of America, 119(2), 1118.

Jia, G., Strange, W., Wu, Y., Collado, J., & Guan, Q. (2006). Perception and production of English vowels by Mandarin speakers: Age-related differences vary with amount of L2 exposurea). The Journal of the Acoustical Society of America, 119(2), 1118-1130.

Jusczyk, P. W. (2000). The discovery of spoken language. MIT press.

Jusczyk, P. W., & Hohne, E. A. (1997). Infants' memory for spoken words. Science, 277(5334), 1984-1986.

Khalil, A. M. (1996). A contrastive grammar of English and Arabic. Al− Isra Press.

Kuhl, P. K. (1983). Perception of auditory equivalence classes for speech in early infancy. Infant Behaviour and Development, 6(2-3), 263–285.

Kuhl, P. K. (2000). A new view of language acquisition. Proceedings of the National Academy of Sciences of the United States of America, 97(22), 11850–7.

Kuhl, P. K. (2004). Early language acquisition: cracking the speech code. Nature reviews neuroscience, 5(11), 831-843.

Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2008). Phonetic learning as a pathway to language: new data and native language magnet theory expanded (NLM-e). Philosophical Transactions of the Royal Society B: Biological Sciences, 363(1493), 979-1000.

Ladefoged, P. (1996). Elements of Acoustic Phonetics (second edition). University of Chicago Press.

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(02), 227–247.

Lehn, W., & Slager, W. R. (1959). A contrastive study of Egyptian Arabic and American English: the segmental phonemes. Language Learning, 9(1-2), 25-33.

Lengeris, A. (2009). Individual differences in second-language vowel learning. PhD thesis. Submitted to UCL.

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. The Journal of the Acoustical Society of America, 128(6), 3757–68. doi:10.1121/1.3506351

Lenneberg, E. H., Chomsky, N., & Marx, O. (1967). Biological foundations of language (Vol. 68). New York: Wiley.

Levy, E. S., & Strange, W. (2008). Perception of French vowels by American English adults with and without French language experience. Journal of Phonetics, 36(1), 141–157.

Li, F., Munson, B., Edwards, J., Yoneyama, K., & Hall, K. (2011). Language specificity in the perception of voiceless sibilant fricatives in Japanese and English: implications for cross-language differences in speech-sound development. The Journal of the Acoustical Society of America, 129(2), 999–1011.

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

Lim, S.-joo, & Holt, L. L. (2011). Learning foreign sounds in an alien world: videogame training improves non-native speech categorization. Cognitive science, 35(7), 1390-405.

Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English/r/and/l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. The Journal of the Acoustical Society of America, 94(3), 1242-1255.

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

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

Macdonald, R. (2011). A comparison of techniques used to perceptually train difficult foreign language contrasts. Edinburgh University Linguistics and English Language Postgraduate Conference, Edinburgh, 11th May 2011.

MacKay, I. R., Flege, J. E., Piske, T., & Schirru, C. (2001). Category restructuring during second-language speech acquisition. The Journal of the Acoustical Society of America, 110(1), 516-528.

Massaro, D. W., & Cohen, M. M. (1995). Perceiving talking faces. Current Directions in Psychological Science, 104-109.

Massaro, D. W., & Light, J. (2003, September). Read my tongue movements: bimodal learning to perceive and produce non-native speech/r/and/l/. In INTERSPEECH.

Massaro, D. W., & Light, J. (2004). Using visible speech to train perception and production of speech for individuals with hearing loss. Journal of Speech, Language, and Hearing Research, 47(2), 304-320.

Massaro, D. W., Bigler, S., Chen, T. H., Perlman, M., & Ouni, S. (2008, September). Pronunciation training: the role of eye and ear. In Interspeech (pp. 2623-2626).

Massaro, D., Cohen, M. M., Meyer, H., Stribling, T., Sterling, C., & Vanderhyden, S. (2011). Integration of facial and newly learned visual cues in speech perception. The American journal of psychology, 124(3), 341-354.

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. Language and Cognitive Processes, 27(7-8), 953-978.

McAllister, R., Flege, J. E., & Piske, T. (2002). The influence of L1 on the acquisition of Swedish quantity by native speakers of Spanish, English and Estonian. Journal of Phonetics, 30(2), 229–258.

McCarthy, K. M., Evans, B. G., & Mahon, M. (2013). Acquiring a second language in an immigrant community: The production of Sylheti and English stops and vowels by London-Bengali speakers. Journal of Phonetics, 41(5), 344–358.

Meador, D., Flege, J. E., & MacKay, I. R. (2000). Factors affecting the recognition of words in a second language. Bilingualism: Language and Cognition, 3(01), 55-67.

Miller, J. L., & Eimas, P. D. (1983). Studies on the categorization of speech by infants. Cognition, 13(2), 135-165.

Mitterer, H., & Ernestus, M. (2008). The link between speech perception and production is phonological and abstract: evidence from the shadowing task. Cognition, 109(1), 168–73.

Morley, J. (1991). The pronunciation component in teaching English to speakers of other languages. TESOL Quarterly, 25(1), 51-74.

Munro, M. J., Flege, J. E., & MacKay, I. R. (1996). The effects of age of second language learning on the production of English vowels. Applied Psycholinguistics, 17(03), 313-334.

Munson, B., Edwards, J., & Hall, K. (2014). Language specificity in the perception of voiceless sibilant fricatives in Japanese and English : Implications for cross-language differences in speech-sound development, 129 (February 2011), 999–1011.

Neri, A., Cucchiarini, C., & Strik, H. (2006). "ASR-based corrective feedback on pronunciation: does it really work?," Proceedings of Interspeech 2006, Pittsburgh, PA, 1982-1985.

Neri, A., Cucchiarini, C., & Strik, H. (2008). The effectiveness of computer-based speech corrective feedback for improving segmental quality in L2 Dutch. ReCALL, 20(02), 225-243.

Neri, A., Cucchiarini, C., Strik, H., & Boves, L. (2002). The pedagogy-technology interface in computer assisted pronunciation training. Computer Assisted Language Learning, 15(5), 441-467.

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.

Newman, D. (2002). The phonetic status of Arabic within the world's languages: the uniqueness of the lughat al-daad. Antwerp papers in linguistics. 100, 65-75.

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 : JSLHR, 50(6), 1496–509.

Oh, G. E., Guion-Anderson, S., Aoyama, K., Flege, J. E., Akahane-Yamada, R., & Yamada, T. (2011). A one-year longitudinal study of English and Japanese vowel production by Japanese adults and children in an English-speaking setting. Journal of phonetics, 39(2), 156-167.

Ortega-Llebaria, M., Faulkner, A., & Hazan, V. (2001). Auditory-visual L2 speech perception: Effects of visual cues and acoustic-phonetic context for Spanish learners of English.

Ouni, S., Cohen, M. M., Ishak, H., & Massaro, D. W. (2006). Visual contribution to speech perception: measuring the intelligibility of animated talking heads. EURASIP Journal on Audio, Speech, and Music Processing, 2007.

Piske, T., MacKay, I. R., & Flege, J. E. (2001). Factors affecting degree of foreign accent in an L2: A review. Journal of phonetics, 29(2), 191-215.

Podlipský, V. J., Skarnitzl, R., & Volín, J. (2009). High front vowels in Czech: A contrast in quantity or quality? In INTERSPEECH (pp. 132-135).

Polka, L., & Bohn, O. S. (1996). A cross-language comparison of vowel perception in English-learning and German-learning infants. The Journal of the Acoustical Society of America, 100(1), 577-592.

Polka, L., & Werker, J. F. (1994). Developmental changes in perception of nonnative vowel contrasts. Journal of Experimental Psychology: Human Perception and Performance, 20(2), 421.

Pruitt, J. S., Jenkins, J. J., & Strange, W. (2006). Training the perception of Hindi dental and retroflex stops by native speakers of American English and Japanese. The Journal of the Acoustical Society of America, 119(3), 1684-1696.

Rizzolatti, G., & Arbib, M. A. (1998). Language within our grasp. Trends in neurosciences, 21(5), 188-194.

Rothauser, E. H., Chapman, W. D., Guttman, N., Nordby, K. S., Silbiger, H. R., Urbanek, G. E., & Weinstock, M. (1969). IEEE recommended practice for speech quality measurements. *IEEE Trans. Audio Electroacoust*, *17*(3), 225-246.

Schmidt, a M., & Flege, J. E. (1995). Effects of speaking rate changes on native and nonnative speech production. Phonetica, 52(1), 41–54.

Schwartz, J. L., Basirat, A., Ménard, L., & Sato, M. (2012). The Perception-for-Action-Control Theory (PACT): A perceptuo-motor theory of speech perception. Journal of Neurolinguistics, 25(5), 336-354.

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.

Sheldon, A. M. Y., & Strange, W. (1982). The acquisition of Ixl and / I / by Japanese learners of English : Evidence that speech production can precede speech perception, 243–261.

Shiller, D. M., Sato, M., Gracco, V. L., & Baum, S. R. (2009). Perceptual recalibration of speech sounds following speech motor learning. The Journal of the Acoustical Society of America, 125(2), 1103–13.

Siyyari, M. (2005). A Comparative Study of the Effect of Implicit and Delayed , Explicit Focus on Form on Iranian EFL Learners ' Accuracy of Oral Production . Science and Technology Department of Foreign Languages. Unpublished Master's Thesis September), 1-178.

Skoruppa, K., Pons, F., Christophe, A., Bosch, L., Dupoux, E., Sebastián-Gallés, N., & Peperkamp, S. (2009). Language-specific stress perception by 9-month-old French and Spanish infants. Developmental Science, 12(6), 914-919.

So, C. K., & Best, C. T. (2010). Cross-language Perception of Non-native Tonal Contrasts: Effects of Native Phonological and Phonetic Influences. Language and Speech, 53(2), 273–293.

Srinivasan, R. J., & Massaro, D. W. (2003). Perceiving prosody from the face and voice: Distinguishing statements from echoic questions in English. Language and Speech, 46(1), 1-22.

Stacey, P. C., & Summerfield, A. Q. (2008). Phoneme-Based Training Strategies in Improving the Perception of Spectrally Distorted Speech. Hearing Research, 51(April), 526–538.

Stevens, G. (1999). Age at immigration and second language proficiency among foreign-born adults. Language in Society, 28(04), 555-578.

Strange, W., Akahane-Yamada, R., Kubo, R., Trent, S. a., Nishi, K., & Jenkins, J. J. (1998). Perceptual assimilation of American English vowels by Japanese listeners. Journal of Phonetics, 26(4), 311-344.

Strange, W., Hisagi, M., Akahane-Yamada, R., & Kubo, R. (2011). Cross-language perceptual similarity predicts categorial discrimination of American vowels by naïve Japanese listeners. The Journal of the Acoustical Society of America, 130(4), EL226–31.

Strange, W., Weber, A., Levy, E. S., Shafiro, V., Hisagi, M., & Nishi, K. (2007). Acoustic variability within and across German, French, and American English vowels: phonetic context effects. The Journal of the Acoustical Society of America, 122(2), 1111–29.

Sueyoshi, A., & Hardison, D. M. (2005). The role of gestures and facial cues in second language listening comprehension. Language Learning, 55(4), 661-699.

Tahta, S., Wood, M., & Loewenthal, K. (1981). Age changes in the ability to replicate foreign pronunciation and intonation. Language and Speech, 24(4), 363-372.

Tajima, K., Kato, H., Rothwell, A., Akahane-Yamada, R., & Munhall, K. G. (2008). Training English listeners to perceive phonemic length contrasts in Japanesea). The Journal of the Acoustical Society of America, 123(1), 397-413.

Terrace, H. S. (1963). Discrimination learning with and without "ERRORS" 1. Journal of the experimental analysis of behaviour, 6(1), 1-27.

Tsukada, K. (2011). The perception of Arabic and Japanese short and long vowels by native speakers of Arabic, Japanese, and Persian. The Journal of the Acoustical Society of America, 129(2), 989-998.

Tsukada, K., Birdsong, D., Bialystok, E., Mack, M., Sung, H., & Flege, J. (2005). A developmental study of English vowel production and perception by native Korean adults and children. Journal of Phonetics, 33(3), 263-290.

Van Dommelen, W. A., & Hazan, V. (2010). Perception of English consonants in noise by native and Norwegian listeners. Speech Communication, 52(11), 968-979.

Van Engen, K. J., & Bradlow, A. R. (2007). Sentence recognition in native- and foreign-language multi-talker background noise. The Journal of the Acoustical Society of America, 121(1), 519.

Vroomen, J., & Baart, M. (2009). Phonetic recalibration only occurs in speech mode. Cognition, 110(2), 254–259.

Wade, T., & Holt, L. L. (2005). Perceptual effects of preceding nonspeech rate on temporal properties of speech categories. Perception & psychophysics, 67(6), 939-50. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16396003

Wang, X., & Munro, M. J. (2004). Computer-based training for learning English vowel contrasts. System, 32(4), 539–552.

Wang, Y., Jongman, A., & Sereno, J. A. (2003). Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training. The Journal of the Acoustical Society of America, 113(2), 1033-1043.

Watson, J. C. (2007). The phonology and morphology of Arabic. Oxford university press.

Werker, J. F., & Curtin, S. (2005). PRIMIR: A developmental framework of infant speech processing. Language Learning and Development, 1(2), 197-234.

Werker, J. F., & Lalonde, C. E. (1988). Cross-language speech perception: initial capabilities and developmental change. Developmental psychology, 24(5), 672.

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. Infant behaviour and development, 7(1), 49-63.

Wik, P. (2011). The Virtual Language Teacher: Models and applications for language learning using embodied conversational agents. Doctoral thesis, Stockholm, Sweden.

Wilson, S. M., Saygin, A. P., Sereno, M. I., & Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. Nature neuroscience, 7(7), 701–2.

Yamada, R. A. (1995). Age and acquisition of second language speech sounds: Perception of American English/r/and/l/by native speakers of Japanese. Speech perception and linguistic experience: Issues in cross-language research, 305-320.

Yeni-Komshian, G. H., Flege, J. E., & Liu, S. (2000). Pronunciation proficiency in the first and second languages of Korean-English bilinguals. Bilingualism Language and Cognition, 3(2), 131-149.

Ylinen, S., Uther, M., Latvala, A., Vepsäläinen, S., Iverson, P., Akahane-Yamada, R., & Näätänen, R. (2010). Training the brain to weight speech cues differently: a study of Finnish second-language users of English. Journal of cognitive neuroscience, 22(6), 1319–32.

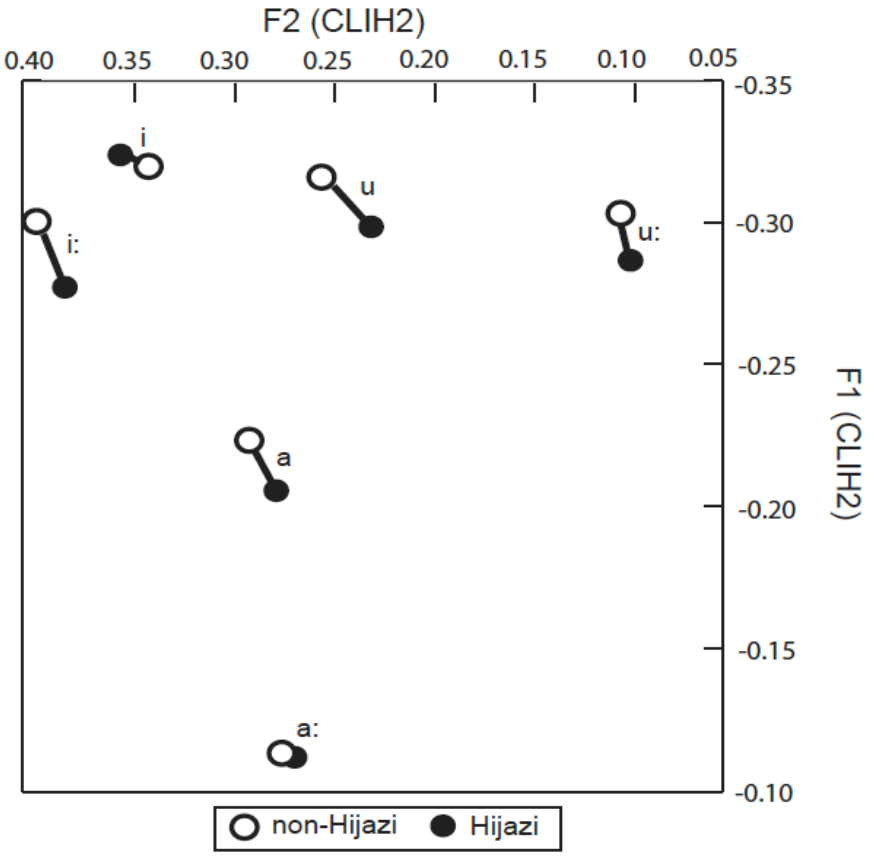
Zhou, X., Woo, J., Stone, M., Prince, J. L., & Espy-Wilson, C. Y. (2013). Improved vocal tract reconstruction and modeling using an image super-resolution technique. The Journal of the Acoustical Society of America, 133(6), EL439.

Appendices

Appendix 1: Arabic consonant phonemes Adapted from (Khalil, 1999)

	labial	Labio-dental	Inter-dental	Alvo-dental	Dental	Alveolar	Post-alveolar	palatal	Velar	Uvular	Pharyngeal	Glottal
Stops	b				t,d, tˤ,dˤ				k	q		`ʔ
Fricatives		f	θ ð ðˤ		s,z,sˤ		ʃ		x, ɣ		ħ ˤ	h
Affricates							dʒ					
Nasals	m					n						
Lateral						l						
Trill						r						
glides	w						j					

Appendix 2: Vowel space produced by Saudi speakers (pilot study



Appendix 3: Confusion matrix showing the percent correct of the vowel intelligibility for L2 learners who were tested in Saudi Arabia at pre-test

	response															
		bait	bart	bat	beat	bert	bet	bit	bite	boat	boot	bot	bought	bout	but	Total
	bait	56	0	2	0	6	13	0	23	0	0	0	0	0	0	100
	bart	1	86	0	0	3	0	0	0	0	0	3	3	0	3	100
	bat	0	3	68	0	11	0	0	0	0	0	0	1	0	17	100
	beat	0	0	1	81	1	8	7	2	0	0	0	0	0	0	100
	bert	1	1	0	9	88	1	0	0	0	0	0	0	0	0	100
	bet	0	0	2	18	1	58	20	1	0	0	0	0	0	0	100
	bit	0	0	4	0	0	50	22	22	0	0	0	0	0	1	100
	bite	2	1	0	0	0	7	3	87	0	0	0	0	0	0	100
	boat	0	0	0	0	0	0	0	0	50	11	10	29	0	0	100
	boot	0	0	0	0	0	0	0	0	17	58	2	21	1	1	100
	bot	0	0	0	0	0	0	0	0	6	14	46	9	1	24	100
	bought	0	1	0	0	0	0	0	0	30	0	16	40	13	0	100
	bout	1	1	0	0	0	0	0	0	13	11	10	3	56	4	100
stimulus	but	0	0	23	0	0	0	0	0	0	0	0	0	0	77	100

Appendix 4: Confusion matrix showing the percentage correct of the vowel intelligibility for L2 learners who were tested in Saudi Arabia at the post-test

	response															
		bait	bart	bat	beat	bert	bet	bit	bite	boat	boot	bot	bought	bout	but	Total
	bait	94	0	0	0	0	2	0	3	0	0	0	0	0	0	100
	bart	0	78	3	0	1	0	0	0	0	1	0	9	2	6	100
	bat	1	4	77	0	4	0	1	0	0	0	0	0	0	12	100
	beat	2	0	0	86	1	9	2	0	0	0	0	0	0	0	100
	bert	0	0	3	0	81	13	0	0	0	0	0	0	1	1	100
	bet	0	0	13	0	2	83	1	0	0	0	0	0	0	0	100
	bit	0	0	2	13	0	41	43	0	0	0	0	0	0	0	100
	bite	1	1	0	0	0	0	0	97	0	0	0	0	0	1	100
	boat	0	0	0	0	0	0	0	0	66	0	2	24	4	3	100
	boot	0	0	0	0	0	0	0	0	8	82	1	9	0	0	100
	bot	0	0	1	0	0	0	0	0	3	0	51	16	0	29	100
	bought	0	0	0	0	0	0	0	0	19	1	7	49	23	1	100
	bout	0	7	0	0	0	0	0	0	1	0	1	8	79	4	100
stimulus	but	0	3	20	0	1	0	0	0	0	0	1	0	1	73	100

Appendix 5: Confusion matrix showing the percentage correct for the vowel intelligibility for L2 learners who were tested in London at the pre-test.

		response													
		bait	bart	bat	beat	bert	bet	bit	bite	boat	boot	bot	bought	bout	but	Total
	bait	74	1	1	0	1	7	4	14	0	0	0	0	0	0	160
	bart	7	66	0	0	3	0	1	0	0	0	1	7	4	11	160
	bat	1	8	72	0	4	5	0	0	0	0	1	0	0	10	160
	beat	11	0	0	79	0	0	9	0	1	0	0	0	0	0	160
	bert	0	3	3	3	89	0	1	0	1	0	0	0	1	1	160
	bet	2	1	8	1	1	48	40	0	0	0	0	0	0	1	160
	bit	0	0	4	5	0	36	42	14	0	0	0	0	0	0	160
	bite	1	1	0	6	0	3	4	87	0	0	0	0	0	0	160
	boat	0	0	0	0	1	0	0	0	76	0	3	1	15	5	160
	boot	0	0	0	0	0	0	0	0	20	65	6	6	0	3	160
	bot	0	1	1	0	0	0	0	0	9	1	43	8	1	37	160
	bought	0	3	1	0	0	0	0	0	68	0	4	9	14	2	160
	bout	0	0	1	0	1	0	0	1	22	0	4	6	66	1	160
stimulus	but	0	1	6	0	0	0	0	0	1	0	4	2	1	86	160

Appendix 6: Confusion matrix showing the percentage correct for the vowel intelligibility for L2 learners who were tested in London at the post-test

		response														
		bait	bart	bat	beat	bert	bet	bit	bite	boat	boot	bot	bought	bout	but	Total
	bait	70	0	1	0	1	8	3	18	0	0	0	0	0	0	100
	bart	0	63	1	0	2	0	0	0	1	0	6	19	3	6	100
	bat	0	8	76	1	2	8	0	0	0	0	0	0	0	6	100
	beat	9	0	1	79	1	6	4	0	0	1	0	0	0	0	100
	bert	1	9	3	0	78	1	2	0	6	0	0	0	0	1	100
	bet	0	0	13	6	3	69	4	0	1	0	0	0	0	4	100
	bit	0	0	1	10	0	21	61	6	0	0	0	0	0	0	100
	bite	1	1	0	5	0	0	1	91	0	0	0	0	1	0	100
	boat	0	0	0	0	0	0	0	0	74	3	6	1	15	2	100
	boot	0	0	0	0	0	0	0	0	7	83	2	8	1	1	100
	bot	0	1	0	0	0	1	0	0	2	6	48	18	1	25	100
	bought	0	1	0	0	1	0	0	0	31	0	3	35	22	7	100
	bout	0	0	0	0	0	0	0	1	8	0	0	3	83	6	100
stimulus	but	0	3	11	0	0	0	0	0	0	0	3	1	1	83	100