Improving the Perception of English Vowels by Arabic-speaking Learners in Saudi Arabia

Sarah Jameel M Alghabban

The thesis submitted for the
degree of Doctor of Philosophy (Linguistics)

School of Education, Communication, and Language Sciences
Newcastle University

January 2024

Abstract

High variability (HV) training method has been found to be effective in improving the perception of English vowels. Its success is derived from several factors, including exposing learners to multiple speakers and phonetic contexts, the provision of immediate feedback after trials, and the utilisation of real examples of natural language. However, current HV training studies mainly focus on using a single first language (L1) English variety in their methodology, often Southern Standard British English (SSBE). The potential advantages of including multiple English accents in the training process have so far been overlooked; this is particularly poignant when it comes to second language (L2) varieties, which are typically avoided. With this in mind, the current study examined whether accent variability aids the perception of English vowels by L2 learners in the same way other aspects of variability have proven beneficial and whether this advantage holds when L2 (or so-called "non-native") varieties are used.

Data were collected from 112 Saudi Arabian novice learners studying English in a foreign language context. Participants were randomly assigned to one of three training groups: Group A included 38 participants who were trained using the HV paradigm with a single L1 accent (SSBE); Group B included 41 participants who were trained using a variety of L1 Standard English accents (SSBE, American, Australian); and Group C included 38 participants who were trained with two L1 varieties (SSBE, American) and one L2 variety (Saudi English). Each training consisted of sixteen 40-minute sessions carried out over a period of three months. To my knowledge, this is the longest training study of its kind, with the highest number of participants. The training method integrated aspects that have been empirically demonstrated to improve learners' perception of vowels; these included identification, auditory discrimination, and category discrimination tasks alongside a production task. Participants completed a pre-test (before the training), a mid-test (after 8 training sessions), and a post-test (after 16 training sessions) to assess the effectiveness of the training. Furthermore, they completed two generalisation tests, which provided novel stimuli not previously used in training or the pre-/mid-/post-tests. The first generalisation test featured new words pronounced by new Saudi and SSBE speakers, while the second test used these same words but spoken with novel accents (Indian and Chinese English). The tasks in all tests (pre-/mid-/post-tests and generalisation tests) were identical, consisting of identification, auditory discrimination, and category discrimination, all of which were performed without any feedback.

The findings revealed that groups A, B, and C significantly improved on all three perceptual tasks, as demonstrated not only by their mid-and post-test results but also by their results on both generalisation tests. Crucially, all three groups improved equally, presenting a solid case for using L2 as well as L1 varieties in phonetics training, given that exposure to different accents was not problematic for beginners. This mimics real-world variability and promotes social justice in the way we portray the role of varieties in the classroom. Moreover, the data showed that the training programme successfully addressed the three perceptual tasks (identification, auditory discrimination, and category discrimination) despite the varying levels of difficulty associated with each task. The data also showed significant improvement in vowel accuracy across the three perceptual tests as time progressed, with the post-test—after 16 training sessions—showing more improvement compared to the mid-test that took place after 8 sessions. Improvement in vowel performance was retained across all tasks and generalisation tests. Given these results, the HV training method developed in this study has the potential to enhance English as a Foreign Language (EFL) teaching by supporting learners to recognise and discriminate challenging L2 sounds. This tool should be valuable in environments with limited resources and for learners who primarily encounter the language in educational settings. Through exposure to high variability stimuli, including different accents, speakers, and phonetic instances, learners are prepared to engage in everyday conversations more confidently.

Dedication

This work is affectionately dedicated to my beloved family in remembrance of my departed father and elder brother, and with a deep love for my mother, husband, siblings, and daughter.

Acknowledgements

I would like to start by expressing my sincere appreciation to Allah for the unwavering support and abundant blessings which have played a pivotal role in completing this research endeavour. I am filled with a profound sense of accomplishment for completing this work amidst the most arduous circumstances I have encountered. I am convinced that overcoming these obstacles has strengthened my commitment and drive to succeed. I would like to convey my gratitude to a number of people whose invaluable contributions were essential in the successful completion of this work.

I am deeply grateful to my supervisors, professors Ghada Khattab and Jalal Al-Tamimi, for their constant encouragement and guidance. Their presence as outstanding experts throughout this journey was indescribable. Their dedication, patience, insightful advice, meticulous attention to detail, insistence on clarity, and eagerness to aid even in the last moments have all played a substantial role in my growth as a researcher and were pivotal in shaping this research. Every moment of their assistance had a profound effect on my journey, turning it into a truly unforgettable experience. I am deeply grateful for their support and belief in my abilities during difficult personal circumstances. Including them in my academic journey has brought me immense delight, satisfaction, and confidence.

Additionally, I sincerely appreciate the esteemed panel members, Dr Damien Hall and Dr Müge Satar— I am deeply grateful to them for their constructive feedback, invaluable recommendations, and inspiration. My appreciation also goes to all of the participants who provided the stimuli for this study, as well as the learners who enthusiastically participated in a lengthy training period. Every participant's dedication and diligence were vital to the successful conclusion of this thesis.

I am deeply blessed to have fulfilled my parents' lifelong dream, a vision deeply held by my beloved lost father, Jameel, and my valued mother, Khadijah. They wished for me to pursue knowledge and finish my studies with passion. This accomplishment is not solely mine; it is an acknowledgement of my parents' continuous support, care, and provision since childhood, leading me to this point of success. Their sacrifices, advice, and encouragement shaped who I am and powered my journey, making this work as much theirs as mine. Indeed, it is difficult to find words that adequately express my gratitude to my parents. A special thanks goes to my lovely husband, Mohammed, whose confidence in my accomplishments has always been a tremendous source of strength and pleasure. His continuous encouragement throughout our challenging moments and constant presence by my side throughout this long journey has been invaluable. I consider myself very fortunate to have him in my life and look forward to embarking on new adventures with him for the rest of our lives. Furthermore, I am genuinely thankful to my brothers and sisters for their unchanging love and care, which has been a constant source of strength. Their reassurance, comprehension during my absence, or simply being present to listen has been invaluable. The sense of belonging and emotional support they have offered throughout this journey has proven to be highly advantageous in maintaining my focus and motivation. I am also profoundly grateful to my little one, whose smiles and laughter have provided me endless joy and inspiration throughout my academic pursuits, constantly reminding me of our world's beauty and marvels.

Declaration

This thesis does not include any content that has been acknowledged for completing any other academic qualification at any university except for a Doctor of Philosophy degree at Newcastle University. This thesis is entirely devoid of any content that has been previously published or authored by any other person.

I consent for this copy of my thesis to be photocopied and loaned when it is deposited in the university library, as well as for the title and summary to be made available to outside organisations.

SIGNED: Sarah Date: 15. 01. 2024

Table of Contents

Abbreviations

Abbreviation	Full term
AmE	American English
AusE	Australian English
SSBE	Standard Southern British English
SA	Saudi Arabic
HA	Hijazi Arabic
MSA	Modern Standard Arabic
CA	Classical Arabic
HV-S	A Training group heard the SSBE variety
HV-M1	A Training group heard a variety of L1 Standard English accents (SSBE, AmE, AusE)
HV-M2	A Training group heard two L1 varieties (SSBE, AmE) and one L2 variety (SA)
EFL	English as a Foreign Language
ENL	English as a Native Language
ESL	English as a Second Language
F0	Fundamental Frequency First
F1	First Formant Frequency
F2	Second Formant Frequency
F3	Third Formant Frequency
FL	Foreign Language
L1	First Language
L2	Second Language
NE	Native English
NNs	Non-Native Speakers
Ns	Native Speakers
NLM	Native Language Magnet Theory
PAM	Perceptual Assimilation Model
SLM	Speech Learning Model
L2LP	Second Language Linguistic Perception
SC	Single Category
TC	Two Categories
CG	Category Goodness

UC	Uncategorised-Categorised
UU	Uncategorisable
IPA	International Phonetic Alphabet
VISC	Vowel Inherent Spectral Change
HVPT	High Variability Phonetic Training
HV	High Variability
ID	Identification
DIS	Discrimination
AD	Auditory Discrimination
CD	Category Discrimination
Gen1	First generalisation test
Gen2	Second generalisation test
AOA	Age of Arrival
CPH	Critical Period Hypothesis
UR	Underlying Representation
SR	Surface Representation
High Proficiency	HP
Low Proficiency	LP
Generalised Linear Mixed Model	GLMM
Estimated Marginal Means	EMMeans
Confidence Intervals	CIs

List of Figures

List of Tables

Chapter 1 Introduction

The introduction chapter aims to provide an overview of the study background by demonstrating the emergence of English as a worldwide language with diverse accents and introducing essential terms, including native-likeness, target-likeness, and intelligibility. The chapter addresses how segmental errors can impair intelligibility. In addition, it highlights the role of phonetic training in supporting second language (L2) learners to improve their perception and production[1] of challenging English sounds. The chapter also highlights the study's significance, the problem statement, and the study's contribution. Finally, the research questions and the structure of the thesis are presented.

1.1 Research background

1.1.1 The Globalisation of English

English is a global language used by L1 and L2 speakers as a medium of intercultural communication. With the growth of English, it has become spoken with a diversity of accents by speakers from various regions of the world. Noting the emergence of multiple varieties of English with distinct accents, speech patterns, and lexicons, researchers like Kachru (1991) have emphasised the importance of recognising the existence of multiple varieties of English, collectively referred to as 'Englishes,' to acknowledge the diversity of its use. While there is no single pathway for the emergence of a new variety of English, it is generally a process of adaptation. A group of speakers takes an existing form of English and modifies it to meet their specific communicative needs, aligning linguistic features with local social and cultural practices (see also Kirkpatrick (2007)).

Kachru (1992) advised the **Three Circle Model**, in which various types of English are shown as three overlapping circles: the Inner Circle, the Outer Circle, and the Expanding Circle. The Inner Circle comprises the core English-speaking nations, primarily associated with L1 speakers from the United Kingdom, the United States,

[1] The terms 'perception' and 'production' are similar to the terms 'recognition' and 'pronunciation' used in other disciplines like teaching and applied linguistics.

Australia, Canada, and New Zealand. It illustrates the conventional foundations of English, where the language became the dominant first language through colonisation and migration. This is particularly evident in ex-colonial regions like Australia, Canada, and New Zealand, where British colonial influence established English as the native language for much of the population. Although other languages are spoken, English remains the predominant official language. The Inner Circle is considered "norm-providing," as it sets global standards for how English should be perceived and used. **The Outer Circle** includes India, Singapore, the Philippines, Ghana, Malaysia, and South Africa, which were part of the former British Empire. English has been institutionalised as an official language and taught as a second language in these settings. Since New Englishes establish their norms independent from the Inner Circle, the Outer Circle is considered "norm-developing". **The Expanding Circle** shows nations with neither a colonial history nor official English legislation, such as Saudi Arabia, China, Egypt, Japan, and Brazil. In these settings, English is taught as a foreign language (FL), and it is "norm-dependent" because it depends on the Inner Circle for norms. The classification of the Three Circle Model is comparable to that of English as a native language (ENL), English as a second language (ESL), and English as a foreign language (EFL).

Various researchers (e.g., McArthur (1998), Bruthiaux (2003) and Jenkins (2009; 2003)) have criticised the Kachru tripartite model. The following are the most common criticisms:

1- **The model emphasises geographical and historical events above sociolinguistic uses of English**

- The model does not provide an accurate depiction of contemporary English usage. For instance, it does not account for ENL-speaking residents of ESL and EFL regions. Many Anglo-Indians and British expatriates reside in India and Hong Kong. In addition, there are various ENL communities in EFL countries, such as the significant British presence in Spain and the Anglo-Argentine community in South America. Due to growing global mobility, ENL, ESL, and EFL speakers can be found worldwide. As a result of increased worldwide mobility and a higher willingness to gain an English education in an ENL country, there are also an increasing number of international students who are ESL or EFL speakers in ENL countries.

- The emphasis on history and geography obscures the rapid evolution of the importance of the English language in many Expanding Circle regions. English has historically had few international users and minimal internal function in these regions. However, the model fails to underline that this circumstance is rapidly evolving. English in EFL contexts is no longer merely a foreign language, as many individuals work in a globalised business environment and must communicate with non-native (NN) English speakers more frequently than with native (N) speakers. In light of this, it is now more acceptable to refer to these speakers as English as a Language Franca (ELF) speakers rather than EFL speakers.
- The model fails to recognise the use of ELF within and between the three circles.

2- The model focuses excessively on colonial history

- The model does not account for Britain's influence in several countries outside Kachru's Outer Circle. For instance, the British held Egypt from 1882 until the end of the First World War, a considerably more extended period than the American colonisation of the Philippines (Bruthiaux, 2003). Nonetheless, Egypt is included in the Expanding Circle and the Philippines in the Outer Circle. Consequently, colonial history alone is insufficient for understanding the complex sociolinguistic applications of English in the contemporary globe.

3- The model fails to represent the actual function of English in multiethnic and monolingual regions

- The model does not adequately capture the actual function of English in both multiethnic and monolingual regions. Kachru's three-way classification oversimplifies the linguistic diversity in many regions and overlooks the changing role of English over time. For example, Canada and South Africa are highly multilingual, yet the model categorises them as Inner and Outer Circle speakers, respectively, disregarding the important linguistic contributions of groups like French Canadians and Zulus. This highlights that the model cannot fully represent the complex and dynamic linguistic landscapes of such multilingual countries.

4- The model presumes a Monolithic standard

- Jenkins (2009, p. 20) and Bruthiaux (2003, p. 169) highlight the difficulty in utilising the model to define speakers in terms of their English proficiency, as well as the absence of an attempt to discriminate between degrees of communicative ability. This viewpoint is consistent with that of several scholars (e.g., Crystal, 1992; Rampton, 1990; Kramsch, 1997) who have criticised the concepts of 'native speaker, 'mother tongue,' and ideologies of what 'good English' sounds like due to the emergence of new English varieties from around the world.

- With the emergence of ELF among L2 speakers rather than between L1 and L2 speakers, this argument becomes even more noteworthy (Jenkins, 2003). Jenkins opposes the categories of native speakers (Ns) and non-native speakers (NNs) and argues as follows: (1) It is offensive to refer to proficient English speakers as non-native. (2) the vast majority of people are bilingual or multilingual, so it is less logical to assume that monolingualism is the standard of the world; (3) such terms (native, non-native) ignore the lingua role of the language; (4) it suggests an oversimplified view of what constitutes an error in English language use; and (5) it creates difficulties with international English tests such as IELTS/ TOFEL because it implies an irrelevant native standard reference point (American, British).

Though Kachru's model (1986) has its limitations, as previously discussed, it is important to acknowledge the significant positive contributions it has made. Despite criticisms, the model has been instrumental in reshaping the understanding of English's global spread and diverse forms. For instance, the term ‘expanding’ implies a process of continuous growth, emphasising linguistic diversity and offering new research opportunities. The model is used by global Englishes researchers to challenge the influence of native English speakers, raise awareness of variances in English, and address ownership issues (Galloway & Rose, 2015). Furthermore, the model has had a strong influenced on language instruction and applied linguistics, particularly in Asia and Africa (Schneider, 2011), by recognising the different ways English is used globally (Crystal, 2003). Morrow (2004) similarly emphasises the importance of preparing L2 learners to engage with the diverse range of English varieties they are likely to encounter in real-world interactions, rather than focusing on a single accent. Reflecting on this, it is essential that these varieties serve as reliable models for learners, ensuring they are intelligible and free from linguistic errors. This approach cultivates a more adaptable and nuanced understanding of English across different contexts, equipping learners to navigate the complexities of global communication more effectively.

Although L2 teaching has encouraged exposure to multiple accents since the early 1980s, most pedagogical materials continue to emphasise instructors' production and one standard L1 variety (Levis & Moyer, 2014). This restricted exposure may limit L2 learners' ability to comprehend unfamiliar accents in real-world situations (Buck, 2001). In addition, restricting exposure to a single standard variety has led many L2 learners to perceive variations as errors rather than regional variations (Mahboob & Elyas, 2014) or less prestigious varieties (Timmis, 2007). For example, Almegren (2018) found that Saudi learners of English preferred traditional British and American accents over other global English varieties. Moreover, they considered any deviation from these standard accents incorrect (Mahboob & Elysa, 2014). This highlights the critical need to diversify production resources to raise students' awareness of the wide variety of English accents spoken around the globe. Interestingly, research indicates that L1 and L2 English instructors have comparable effects on production instruction. For instance, Levis et al. (2016) investigated the influence of instructors' native languages on student production ratings. The study involved two female PhD candidates from applied linguistics and technology: one native American English speaker and one native Turkish speaker, who taught identical production courses over 7 weeks. Even though students frequently preferred the native teacher, comprehensibility ratings for students taught by both types of teachers were comparable. This suggests that effective production instruction depends more on skilful instruction than on the instructor's accent.

While standard varieties like Received Pronunciation (RP) and General American are often favored in language training due to the belief that they are more intelligible and authentic, this preference primarily arises from familiarity and established conventions rather than any inherent linguistic superiority. The perception of these accents as clearer or easier to understand is largely shaped by learners' frequent exposure to them in educational settings, rather than an objective measure of clarity or ease of understanding. In today's globalised world, where both L1 and L2 speakers are easily accessible, it becomes difficult to justify focusing exclusively on one variety of English for L2 learners. Interlocutors in real-world contexts can be either L1 or L2 English speakers, and L2 learners may encounter a wide range of English accents, including both standard and regional varieties. Therefore, it is essential not to privilege one variety over another. L2 learners should be exposed to multiple English varieties, including standard L1 variants (e.g., American, British) as well as regional and L2 English variants. This approach ensures that learners are better prepared for the linguistic diversity they will encounter and fosters awareness of the broad spectrum of English accents spoken globally. This need for exposure to multiple varieties is especially important in foreign language contexts, where learners often encounter non-native accents more frequently than native ones. In Saudi Arabia, for instance, learners are primarily surrounded by a multilingual environment where various L2-accented Englishes are prevalent, rather than native-standard varieties. They are more likely to encounter non-native English accents from their own region (Saudi-accented English) as well as from other L2 speakers such as Indian and Pakistani English speakers. Additionally, they are often taught by lecturers and interact with peers who speak Saudi-accented English. The argument is not to exclude native standard varieties from training or educational settings, but rather to ensure that learners' exposure is not confined to a single native variety, as often seen in past studies (e.g., Shinohara & Iverson, 2021).

Therefore, this study intends to bridge the gap between academic language instruction and practical language use in the foreign language context of Saudi Arabia by incorporating both standard and non-standard accents. It includes a broader range of accents—Saudi-accented English alongside standard American, British, and Australian varieties—to enhance learners' ability to perceive diverse linguistic inputs. Rather than encouraging the imitation of specific native or non-native accents, the focus is on improving learners' perceptual flexibility and comprehension of English vowels across different accents. Standard accents were included to facilitate comparison with prior studies that often rely on a single native variety. Importantly, the study does not favor specific native or non-native accents but instead prepares learners to navigate a wide spectrum of English accents, offering a more comprehensive approach to language learning. Given the inconsistency in the classification of English speakers in the literature (e.g., N/NN speakers, EFL/ESL), this study adopts the terms L1 and L2 speakers.

1.1.2 Nativeness, Targetness and Intelligibility

Before the 1960s, the **nativeness** principle was the prevailing viewpoint in production research, with native-like performance viewed as a desirable and attainable objective for learners. To acquire speech patterns identical to those of native (L1) speakers, students must acquire an accurate imitation of a specific native speech model (Kang, 2010). It is sometimes argued (e.g., Harmer, 2001) that students should have the opportunity to achieve a native-like accent if they so desire. Although some research indicates that adult L2 learners may occasionally attain native-like speech patterns, this phenomenon appears limited to a very small number of highly motivated individuals and those with exceptional aptitude (Ioup et al., 1994). Since native production and perception of L2 sounds typically occur during childhood (Lenneberg, 1967), Levis (2005) argues that expecting adult L2 learners to establish a native-like accent is unrealistic. Rather than prioritising a native-like accent, a **"target-like accent"** is a more practical production goal that L2 learners should strive for. While Ur (1996) does not explicitly define "target-like production", she underscores the significance of developing a clear and comprehensible production that aligns with the learner's needs and goals rather than solely aiming for a perfect imitation of native speech patterns.

Willis (1996) further suggests that instead of aiming for a native speaker's version, an "internationally accepted version" (p.12) of the target language could be used. This, however, raises the question of how such a version would be defined and whether it would create another standard, which applied linguists argue against. Nonetheless, Willis seems to advocate for a version of the language that reflects the practical and skilled use of English by L2 speakers in global contexts, prioritising communicative effectiveness over the pursuit of native-like proficiency. Additionally, Cook (1999) advocates for the inclusion of proficient L2 speakers in language instruction, recognising them as capable users of the language rather than deficient non-native speakers. In line with this view, audio recordings used in the classroom should feature proficient L2 speakers to reflect the diversity of language use and offer learners realistic models to follow.

Recently, there has been a shift in L2 production pedagogy from focusing on native-like performance to prioritising **intelligibility** (Derwing & Munro, 2015). It is important to note that intelligible speech does not necessarily equate to target-like or native-like production, as it is possible for speech to be intelligible despite production errors (Munro & Derwing, 1995) or strong accents. The updated descriptors in the Common European Framework of Reference (CEFR) underscore this shift by placing intelligibility at the forefront of phonological control. The emphasis on "accent and accuracy rather than intelligibility has hindered the development of production instruction" (Council of Europe 2018, p.134). The concept of intelligibility dates back to 1949 when Abercrombie suggested that learners need only "comfortably intelligible pronunciation" (Abercrombie, 1949, p.3), meaning they can be understood with little or no conscious effort (p.120). Abercrombie's interpretation of intelligibility aligns with Munro and Derwing's (1995a) notion of "comprehensibility," or the ease with which speech is understood. In exploring how listeners manage to understand speakers from various English varieties, Smith and Nelson (1985) developed a three-part system:

- **Intelligibility**: The recognition of words and utterances (i.e., listeners can repeat or transcribe the spoken words).
- **Comprehensibility**: The degree to which the listener understands the meaning of the words (locutionary force).
- **Interpretability**: The listener's understanding of the speaker's intention behind the message (illocutionary force).

For instance, if someone says, "It is hot in here," the message is intelligible if the listener correctly distinguishes "hot" from "hat" and "here" from "there." It is comprehensible if the listener understands the speaker is referring to the room's temperature. The message is interpretable if the listener responds by adjusting the temperature.

Munro and Derwing (1995, 1997) built upon this framework by introducing a three-part system that has profoundly influenced contemporary pronunciation research:

- **Intelligibility**: Like Smith and Nelson (1985), this refers to the listener's ability to understand the words spoken. It focuses on the recognition of utterances, different from overall listening comprehension, which may use top-down processes to fill gaps in unintelligible speech.
- **Comprehensibility**: This relates to the listener's perception of how much effort is required to understand the speaker. Munro and Derwing (1995) argue that two foreign-accented utterances may be equally intelligible but differ in how much mental effort is needed to process them.
  - **Accentedness**: This refers to the listener's perception of how much the speaker's accent deviates from the target norm. They concluded that intelligibility is not necessarily linked to accentedness; a speaker with a strong accent may still be highly intelligible.

While intelligibility generally refers to the ability to perceive speech sounds and utterances (Smith & Nelson, 1985; Munro & Derwing, 1995; Trudgil, 2003; Kachru & Smith, 2008), some researchers (e.g., Hahn, 2004; Gass & Varonis, 1984; Isaacs & Trofimovich, 2012) confuse it with comprehensibility, which relates to the ease of understanding meaning. This inconsistency in defining intelligibility leads to varied measurement methods (Munro & Derwing, 1995; Field, 2005; Kirkpatrick et al., 2008; Pickering, 2006). When intelligibility is defined as the recognition of speech sounds, measures like orthographic transcription (Kang et al., 2018), forced-choice identification tasks (Hayes-Harb et al., 2008), and true/false comprehension questions (Hahn, 2004; Derwing & Munro, 2014) are used. However, when intelligibility is conflated with comprehensibility, subjective ratings are often used (e.g., Win, 1998; Kirkpatrick et al., 2008), which can obscure the distinction between these two concepts. Jenkins (2000) highlights the ongoing debate about intelligibility, stating that it "can mean different things to different people" (p. 70). This thesis adopts the original definition proposed by influential scholars such as Munro & Derwing (1995) and Smith and Nelson (1985), emphasising the perception of speech sounds. While intelligibility and comprehensibility are closely linked, speech can be intelligible (words are recognised) yet still difficult to fully comprehend, requiring additional cognitive effort.

1.1.3 Segmental errors and Intelligibility

Segmental errors refer to errors in individual sounds (vowels, consonants), confusing whether specific words were heard or not. These errors can substantially impact speech perception, leading to reduced intelligibility (Levis, 2018; Schairer, 1992). For example, Bent et al. (2007) examined the relationship between segmental accuracy and the overall intelligibility of sentences produced by Mandarin learners of English. Segment production accuracy for vowels and consonants across different word positions was assessed. The study found that vowel accuracy had the most significant effect on speech. That is, when L2 learners mispronounced vowels, it was more challenging for L1 English listeners to comprehend their speech.

In an analysis of research on intelligibility, Rajadurai (2007) found that numerous studies adopt a one-dimensional perspective on intelligibility, wherein the listener's assessment of the speaker's clarity is the sole focus. However, Rajadurai argues that communication is a dynamic process in which roles often interchange; the listener becomes the speaker and vice versa. Hence, the degree of intelligibility is not solely contingent upon the speaker's production but rather a collective effort involving all interlocutors in the conversation. In other words, increasing both the perception and production of sounds would contribute to overall intelligibility improvement. The speaker ought to strive for intelligible production, while the listener should aim for meticulous perception. Listed below are examples of segmental errors in English and Arabic languages that could impact the overall intelligibility of a conversation:

Language	**Examples of segmental errors**
English	Producing /əʊ/ as [aʊ]: /kaʊtʃ/ (~*couch*) instead of /kəʊtʃ/ (*coach*) Producing /f/ as [θ]: /θriːz/ (~*trees*) instead of /friːz/ (*freeze*) Producing /u/ as [uː]: /puːl/ (~*pool*) instead of /pul/ (*pull*) Producing /iː/ as [i]: /sin/ (~*sin*) instead of /siːn/ (*scene*) Producing /ɔː/ as [ɒ]: /dɒn/ (~*don*) instead of /dɔːn/ (*dawn*)
Arabic	Producing /q/ as [k]: /kalb/ (~*dog*) instead of /qalb/ (*heart*) Producing /i/ as [iː]: /khariːf/ (~*autumn*) instead of /kharif/ (*senile*) Producing /z/ as [s]: /sir/ (~*secre*t) instead of /zirr/ (*button*) Producing /a/ as [iː]: /khabiːr/ (~*expert*) instead of /khabar/ (*news*)

Table 1.1 Examples of segmental errors in Arabic and English languages.

1.1.4 Phonetic training

Previous research has consistently shown that L2 learners struggle to perceive and produce segments that do not exist in their L1 language system (Bohn & Munro, 2007). For example, research indicates that Spanish and Arabic learners of English frequently classify multiple L2 vowel categories into the same category as their L1 (e.g., Iverson & Evans, 2007; Alshangiti, 2015), whereas Japanese learners of English find it challenging to differentiate between the L2 sounds /l/ and /r/ (Shinohara & Iverson, 2018).These obstacles will likely affect intelligibility, highlighting the need for phonetic training to help learners overcome these difficulties.

Phonetic training provides students with the required input and feedback to improve their perception and production of challenging L2 sounds, boosting their intelligibility. According to Iverson et al. (2012, p.15), phonetic training "provides a useful addition to real-word L2 experience". Most of the success observed in phonetic training research has been linked to the exposure of learners to highly variable natural speech material, commonly known as **high variability phonetic training (HVPT)**, **high variability (HV) training** (Iverson et al., 2012; Lively et al., 1994), or **high variability pronunciation training** (Thomson, 2018). The method's effectiveness has been theoretically linked to the input variability, which includes multiple talkers and phonetic contexts (Logan et al., 1991). It has been hypothesised that the training programme ought to provide learners with the necessary skills to effectively handle the wide range of variations encountered in natural speech. This can be achieved by assisting learners in focusing on the relevant parameters that discriminate the challenging L2 consonant sounds (Iverson et al., 2005) and L2 vowel sounds (Iverson et al., 2023**;** Iverson et al., 2012; Rato, 2014). Thus, they shift their focus from specific talker-related traits to the specific properties of the targeted sounds.

Along with the benefits of highly variable input, the effectiveness of the HVPT method has been attributed to its capacity to provide immediate feedback following each trial. Researchers such as McCandliss and colleagues (2002) highlight this continuous feedback as a distinguishing feature of the method's explicit training approach. HVPT also been shown to promote the generalisability of learning and long-term knowledge retention. Lively et al. (1994) found that learners not only applied their improved skills to new, untrained stimuli but also retained these perceptual gains three months after the training ended. This suggests that HVPT helps learners develop more robust and lasting phonetic representations, which are critical for successful L2 acquisition.

Although HVPT primarily focuses on perception, some studies have shown that perceptual training can lead to improvements in production, albeit to a lesser extent than in perception (Rato & Rauber, 2015; Pereira, 2014; Bradlow, 2008; Hardison, 2004; Thomson, 2012). However, the extent to which perception training reliably transfers to production is still debated. For instance, Bradlow et al. (1997) showed a clear transfer of perceptual gains to production, while Hattori & Iverson (2007) found no such improvement in production following perceptual training. This inconsistency in the literature underscores the need for caution when assuming that perception training will reliably transfer to production. In the current study, a combined HV perception and production training approach was used to improve learners' perception of English vowels, building on research indicating that this method supports perceptual gains and is comparable to perception-focused training (Alshangiti, 2015; Wong, 2014). Since the primary goal of this study was to assess perceptual improvement, production outcomes were not measured or analysed. Therefore, any potential transfer from perception to production falls outside the scope of this research, and no assumptions are made regarding production improvements.

Much of the existing research has been centred on evaluating the efficacy of the HVPT method using identification (ID) tasks. In these tasks, participants are tasked to choose the stimulus they have heard from a close set of options followed by trial-by-trial feedback (e.g., Logan et al., 1991). For example, if participants hear the stimulus "met," they are given responses "meat, met, mitt" in which their goal is to correctly label or identify the word just heard. The widespread incorporation of ID tasks in HVPT can be attributed to the influential guidance provided by Logan and Pruitt in 1995. They advocated for the inclusion of identification tasks in perceptual training, asserting that they demonstrate greater generalisability to unfamiliar speakers and novel words. As a result, there has been a noticeable decline in research that centres on discrimination (DIS) tasks.

While previous efforts at discrimination training have demonstrated only modest success, it is crucial to mention that these endeavours did not incorporate highly variable speech into the DIS tasks (Strange & Dittmann, 1984). Some studies have compared the impact of HV phonetic training using both ID and DIS tasks (Carlet & Cebrian, 2019; Shinohara & Iverson, 2018; Nozawa, 2015; Wayland & Li, 2008; Flege, 1995a). These studies generally suggest that the type of task—whether ID or DIS—does not significantly affect overall outcomes, as both tasks have been shown to improve learner performance to a similar degree (Shinohara & Iverson, 2018; Wayland & Li, 2008; Flege, 1995b).

However, Carlet and Cebrian (2019) found that while both tasks contribute to L2 learning, ID tasks may offer an advantage in terms of generalisation. This difference could be attributed to task familiarity, as the ID group was tested using the same identification tasks that were used during training. In other words, the ID group's better generalisation may be due to their familiarity with the task format throughout the training phase, giving them an advantage in applying their learning to new contexts. The absence of a DIS test in the study may have put the DIS group at a disadvantage by limiting their familiarity with the task. Nevertheless, earlier research by Flege (1995) and Carlet (2017), which assessed both ID and DIS training effects using only ID tasks, found no significant differences between the two groups in perceiving English stop consonants. Despite these findings, additional research is required to determine which task yields the most significant improvements in L2 sound perception or to further validate whether both tasks develop similar gains in performance. In addition, considering the prolonged duration of three months for the current training study, it is vital to incorporate a diverse range of perceptual activities. This will not only enrich the learning process but also ensure that participants remain actively involved throughout the training period.

1.2 Significance of the study

In a variety of fields, such as academia, business, and tourism, oral English proficiency is gaining importance in non-English speaking countries. Saudi Arabia, for example, stands out as a country with a great need for proficient English speakers. In recent years, Saudi Arabia has placed significant emphasis on developing advanced English communication skills, influenced by the country's growing international collaborations and projects. Central to this progress is **the Saudi Vision 2030 initiative**, which was introduced in 2016. Its primary objectives are to establish a flourishing community, promote economic development, and build an innovative country. In light of the global trend towards renewable energy and the fluctuation of oil prices, Saudi Arabia is eager to diversify its economic foundation (Saudi Vision 2030, 2023). Therefore, improving the English proficiency of the workforce is essential for attracting international investors and enterprises. In accordance with Saudi Vision 2030, **the General Entertainment Authority** was established to supervise and advance the entertainment industry. Their mission is to provide world-class entertainment options that actively support the nation's goals of supporting a dynamic society and a thriving economy. The advancement of the tourism sector in Saudi Arabia, which includes the establishment of novel tourist destinations and the organisation of global events, necessitates proficient English communication abilities to deliver exceptional customer service and entice visitors from various parts of the globe (Authority General Entertainment, 2023). Possessing an excellent knowledge of both written and spoken English and Arabic is essential for positions at the General Entertainment Authority, as pointed out in their recent job advertisements (Wadifa.com, 2023; Wadhefaplus, 2023). The prominence of the English language in Saudi Arabia also extends to various aspects of employment. Employers across different private sectors, such as industries, hospitals, and hotels, commonly expect applicants to show certain levels of English proficiency (Al-Seghayer, 2023).

Additionally, the **NEOM project**, with its $500 billion investment, will create a technologically advanced megacity designed to attract global businesses, where English proficiency will be essential for effective collaboration. NEOM has launched its Scholarship Programme as part of its mission to empower the next generation. This programme allows secondary school graduates from areas like Tabuk, Alwajh, Duba and Haql to enrol in highly competitive universities within the country and overseas. Students begin with a foundational year at Fahd bin Sultan University in Tabuk. After this, they proceed to specialised courses at renowned institutions abroad (e.g., UK, US) or in Saudi Arabia. To qualify for international scholarships, students must score at least a 6 on the IELTS or its equivalent. For local scholarships, a minimum score of 5 on the IELTS or its equivalent is required (NEOM, 2020). After graduation, these students will shape the future of NEOM by joining its numerous sectors or affiliated businesses and interacting with L1 and L2 English speakers worldwide. Furthermore, Saudi Arabia offers a variety of scholarships to students who wish to study at highly ranked universities abroad through the Custodian of the Two Holy Mosques Scholarship Programme. Saudi Arabia provides scholarships for students to study overseas via the Custodian of the Two Holy Mosques Scholarship Programme. To be eligible, students must secure unconditional offers, which mandate a minimum English proficiency score for their chosen program. This underscores the need for Saudi students to improve their spoken and written English abilities substantially.

Although oral English proficiency increases educational and professional opportunities for Saudi students, less emphasis has been placed on developing the production and perception of English sounds in the classroom (e.g., Al-Nasser, 2015;Al-Shaibani, 2023). This is consistent with existing literature where, in numerous settings beyond Saudi Arabia, production training is often underemphasised. For instance, Olson's (2014) survey revealed that most L2 instructors devote no more than three minutes per class to production instruction. Surprisingly, 77% of these teachers acknowledge that this is insufficient. In another study, Foote et al. (2011) found that ESL teachers devote only 6% of class time to production instruction. Additionally, Huensch (2019) conducted a study investigating the training, classroom practices, and beliefs associated with production instruction. The sample comprised 296 Spanish, German, and French instructors teaching introductory language courses at 28 large public institutions in the United States. The instructors responded to an online survey in which they reported (1) how frequently production is taught, (2) what production activities they employ in the classroom, and (3) whether they have access to, or desire continued training opportunities. 82% of instructors surveyed reported spending less than 15 minutes per week on production instruction. Typical listen-and-repeat techniques were utilised in both classroom and online activities. Although instructors reported having the ability and self-assurance to teach production, the majority lacked training or access to it and desired additional training in production instruction.

1.3 Statement of problem

Accurately perceiving and producing English vowels is crucial for effective communication, particularly in academic and professional contexts (Munro & Derwing, 2006). The difficulty in this area can reduce intelligibility, result in communication breakdowns and misunderstandings when interacting with L1 and L2 English speakers. Despite the importance of oral skills, Saudi Arabic (SA) learners of English (the participants of this study) face challenges in accurately perceiving and producing English vowels (Evans & Alshangiti, 2018; Alshangiti, 2015). This is due to the spare vowel systems in their L1 in comparison to the denser and more complex systems found in English varieties. In addition, the phonetic and phonological distinctions between Arabic and English are significant (See Almurashi et al., 2023; Algethami, 2023; Alshangiti, 2015).

Research has consistently shown that effective teaching methods are needed to address these difficulties. For instance, Al-Nasser (2015) found that Saudi English language learners struggle to distinguish between English and Arabic sounds. The study also revealed that many EFL instructors continue to employ outdated instructional strategies, relying heavily on the grammar-translation method. In addition, the teaching methods used are primarily teacher centred. Further research by Alfallaj (2013), Tersta & Novianti (2017), Ababneh (2018), Al-khresheh (2020), Alzamil (2021), Alseadan (2021) and Al-Shaibani (2023) has consistently pointed out the challenges Saudi students experience in areas of oral skills, including production, listening, and speaking. Although there is a growing need to improve the oral communication skills of Saudi learners, this study will specifically focus on improving the perception of English vowels.

One potential approach to address the difficulty of perceiving English vowels is the use of HV training method, which has been successful in improving the perception and production of English vowels (Bradlow et al., 1999; Zhang et al., 2021; Iverson et al., 2023). The technique exposes students to a diversity of English vowel sounds produced by various speakers, thereby improving their ability to perceive and produce vowels. In addition, research has shown that improving learners' ability to perceive vowel sounds can positively impact their ability to produce them accurately (Best & Tyler, 2007). By implementing the technique in settings where exposure to English outside of the classroom is limited, the quality of instruction can be enhanced. This can assist educators and curriculum coordinators in refining and aligning their methods with current L2 research and the evolving demands of their students.

While the study employs the HV training method to improve SA learners' perception of English vowels, it can also apply to learners from other languages with a limited vowel set, such as Japanese and Spanish speakers, among others. Given that these learners may face comparable challenges, the methodologies presented in this study could boost their language learning experience. Furthermore, this study addresses the growing demand for efficient production strategies in foreign language classrooms (e.g., Al-Shaibani, 2023; Huensch, 2019; Thomas, 2018). The diagram presented below provides a clear illustration of the research problem.

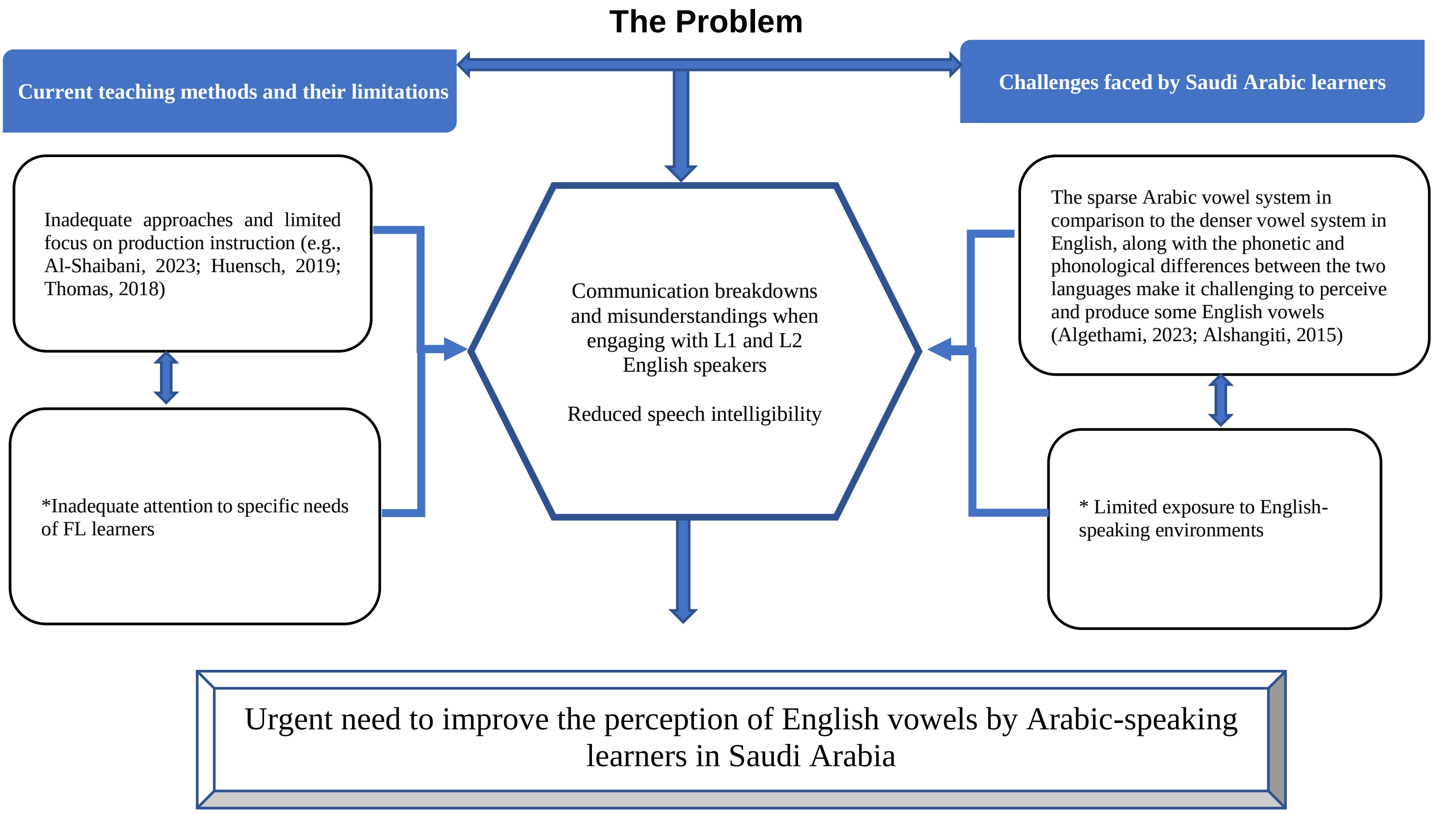

The Problem
Current teaching methods and their limitations
Challenges faced by Saudi Arabic learners
Inadequate approaches and limited focus on production instruction (e.g., Al-Shaibani, 2023; Huensch, 2019; Thomas, 2018)
*Inadequate attention to specific needs of FL learners
Communication breakdowns and misunderstandings when engaging with L1 and L2 English speakers
Reduced speech intelligibility
The sparse Arabic vowel system in comparison to the denser vowel system in English, along with the phonetic and phonological differences between the two languages make it challenging to perceive and produce some English vowels (Algethami, 2023; Alshangiti, 2015)
* Limited exposure to English-speaking environments
Urgent need to improve the perception of English vowels by Arabic-speaking learners in Saudi Arabia

Figure 1.1 Research problem.

1.4 Originality of the study

The HVPT approach has been shown to be successful in improving the perception and production of English sounds (Bradlow et al., 1999; Lively, 1994). However, the existing phonetic training literature has a substantial focus on a single standard model (e.g., Grenon et al., 2019; Iverson et al., 2012). Therefore, the effectiveness and feasibility of using HVPT to improve the perception of English sounds across multiple accents remains uncertain. This means that while the technique has shown promise in improving the perception of sounds, it is not yet clear whether it can effectively enhance the understanding of different English accents. This thesis's main contribution is utilising the HVPT technique with multiple English L1 and L2 varieties to underscore the significance of accent diversity. Another aim of the study is to focus on the efficacy of the tasks used. Some research suggests that ID and DIS training can result in comparable learning outcomes when conducting HVPT (Shinohara & Iverson, 2018)[2]. Other research suggests that while both ID and DIS tasks are effective, the ID task may lead to better generalisation (e.g., Carlet & Cebrian, 2019), with possible effect being due to task familiarity as noted earlier. Accordingly, this study aims to evaluate the use of perceptual tasks—identification (ID), auditory discrimination (AD), and category discrimination (CD)— not only in the training phase but also in the testing phase. Note that throughout this thesis, the term 'auditory discrimination task (AD)' refers to a task involving natural recordings of minimal pair trials (e.g., 'soap, soup, soap,' where two words are the same and one is different). This differs from the typical use of the term, which often implies the use of synthetic or signal-processed stimuli. Using natural recordings allows for the examination of real-world acoustic variability and follows the methodology described by Shinohara and Iverson (2018).

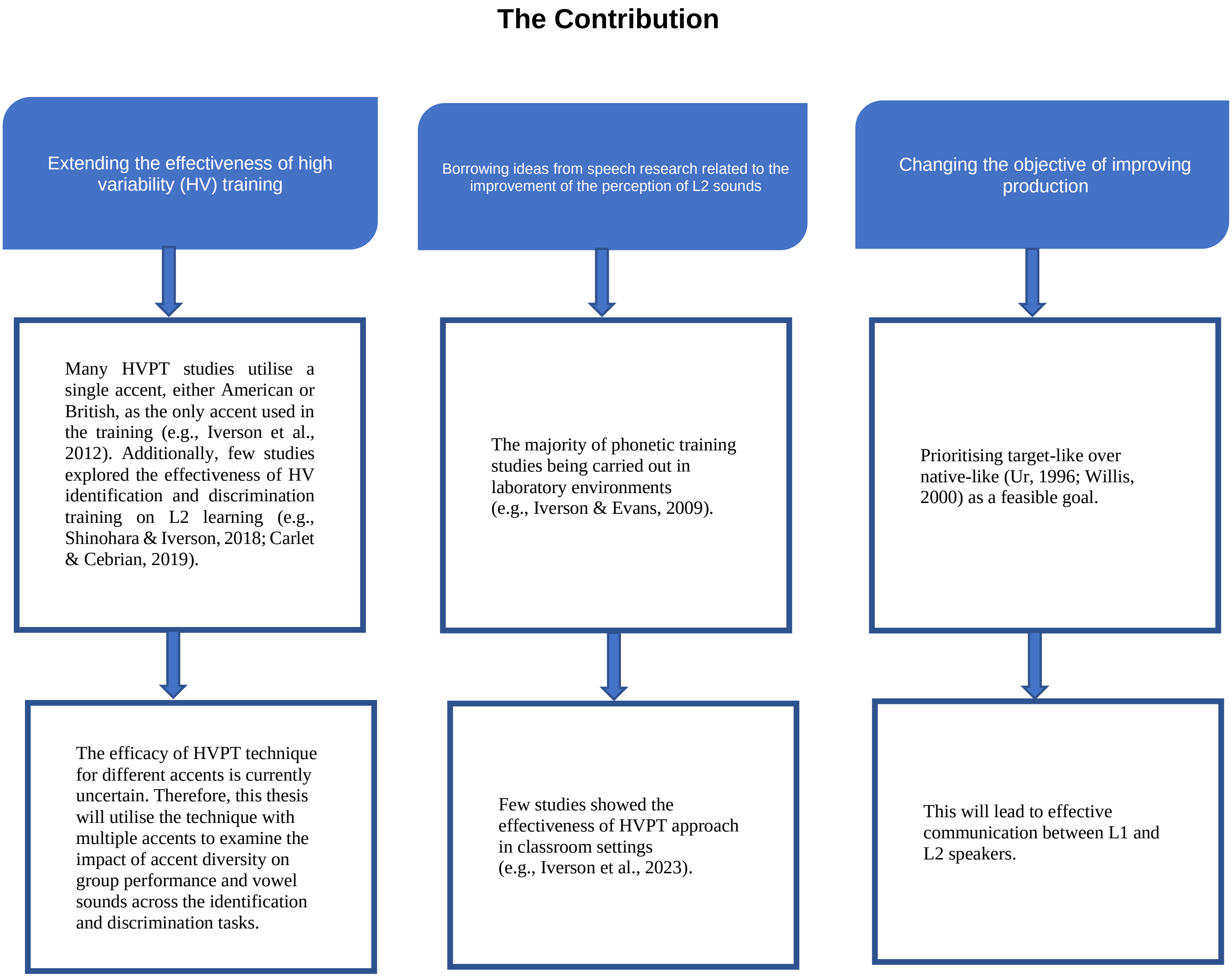
Numerous studies on phonetic training have been restricted to experimental settings (e.g., Iverson & Evans, 2009; Thomson, 2018). This has resulted in a lack of awareness among educators regarding effective strategies to improve students' production and perception of complex L2 sounds. There is a noticeable shortage of

[2] In their study, they employed identification training, which involved a two-alternative force-choice identification task (ID), and discrimination training, which included three tasks: (1) auditory discrimination (AD) with natural recordings, (2) AD using signal-processed stimuli, and (3) category discrimination (CD).

classroom-based research on L2 sound production and perception, with educators favouring traditional teaching methods and providing few opportunities for hands-on or independent learning (Thomson, 2018). A recent study by Iverson et al. (2023) has showed the power of HV perception training, showcasing its effectiveness in enhancing Spanish students' ability to perceive English vowels when taught online. Additionally, Iverson et al. (2012) has demonstrated the advantageous effects of this technique for learners in both immersive and non-immersive environments. Based on these findings, it appears beneficial to integrate research methods on L2 speech into classroom instruction. Thus, the current study addresses the growing need for effective strategies to improve the perception of difficult sounds in classroom-based research. To achieve this, the study combines insights from L2 speech studies and utilises the HV training method to enhance the ability of FL students to distinguish English vowels.

The diversity of English accents spoken around the globe presents a significant challenge for language learners, making it essential to strive for **a target-like accent**. Target-like production is not only desirable but also attainable for learners, as it can enhance their communicative competence and facilitate social integration (Ur, 1996; Willis, 2000). When learners attain target-like accents, they can communicate effectively with L1 and L2 speakers. Remember that a target-like accent differs from a native-like accent (see section 1.1.2). Although this study concentrates on improving the perception of English vowels, it recommends that learners strive for a target-like production as a feasible goal to cultivate a natural and effective communication style that aligns with their target language community's norms and expectations.

Figure 1.2 depicts the innovative contribution of this thesis, which encompasses the integration of multiple accents into the HVPT technique to examine the impact of multiple accents on perceptual tasks and vowel sounds, the implementation of the method in a classroom setting, and the emphasis on prioritising improvement in target-like production rather than native-like production.



The Contribution

Extending the effectiveness of high variability (HV) training

Many HVPT studies utilise a single accent, either American or British, as the only accent used in the training (e.g., Iverson et al., 2012). Additionally, few studies explored the effectiveness of HV identification and discrimination training on L2 learning (e.g., Shinohara & Iverson, 2018; Carlet & Cebrian, 2019).

The efficacy of HVPT technique for different accents is currently uncertain. Therefore, this thesis will utilise the technique with multiple accents to examine the impact of accent diversity on group performance and vowel sounds across the identification and discrimination tasks.

Borrowing ideas from speech research related to the improvement of the perception of L2 sounds

The majority of phonetic training studies being carried out in laboratory environments (e.g., Iverson & Evans, 2009).

Few studies showed the effectiveness of HVPT approach in classroom settings (e.g., Iverson et al., 2023).

Changing the objective of improving production

Prioritising target-like over native-like (Ur, 1996; Willis, 2000) as a feasible goal.

This will lead to effective communication between L1 and L2 speakers.

Figure 1.2 Contribution of the present study.

1.5 Research questions

To address the research problem and contribute to the field of second language production teaching and learning, this study seeks to answer the following main research questions:

RQ1. Does a training program with high variability consisting of a) multiple L1 varieties and one L2 variety benefit FL learners' perception of English vowels to the same extent as a training program with b) multiple L1 varieties or with c) just one L1 variety?

The sub-questions related to this question are as follows:

- For each perceptual task used in this study (ID, AD, and CD)[3], how does each group's response accuracy compare at pre-, mid-, and post-tests?
- For each perceptual task used in this study (ID, AD, and CD), do particular vowels impact the accuracy of responses?

RQ2. Does a training program with high variability consisting of a) multiple L1 varieties and one L2 variety show generalisation effects for FL learners to the same extent as a training program with b) multiple L1 varieties or with c) just one L1 variety?

The sub-sections related to this question are as follows:

- For each task used in this study (ID, AD, and CD), how does each group's response accuracy compare at pre-, post-, and the first generalisation (gen1) tests?
- For each task used in this study (ID, AD, and CD), how does each group's response accuracy compare pre-, post-, and second generalisation (gen2) tests?
- For each generalisation task used in this study (ID, AD, and CD), do certain vowels impact the accuracy of responses?

1.6 The Structure of the Study

This thesis is divided into six chapters as follows:

Chapter 1 offers an introductory overview of the research, emphasising its significance. It proceeds to present the problem statement, followed by an explanation of the study's contribution. Additionally, the chapter lists the research questions that will guide the investigation.

[3] ID refers to the identification task, AD refers to the auditory discrimination task, and CD to the category discrimination task.

Chapter 2 provides an in-depth review of the existing body of research that is relevant to the study. It starts by advocating that L2 learners should not aim for native-like perception and production but rather striving for intelligibility and target-like performance. In addition, it provides an overview of studies about L2 phonetic training, with a particular focus on HV training approach. It also explores how previous exposure to a native language influences perceptual abilities, moving from a broad to a more language-focused perspective. The chapter then proceeds to review L2 models (e.g., speech learning model, perceptual assimilation model of second language). It then examines the differences between vowel systems in the Saudi Arabic variety and those in L1 English varieties, including British, Australian, and American. The chapter concludes by exploring potential difficulties Arabic learners face with shared and unshared vowels, drawing on phonetic and phonological research from various studies on Saudi Arabic and English.

Chapter 3 details the methodologies applied in the study. The chapter provides a description of the participants who were recruited for the study, the methodology employed for delivering the training, and the plan that was implemented for the training. Furthermore, it presents a detailed explanation of the design for perceptual and production training, the methodologies implemented for data collection, the stimuli utilised, and the method employed for data analysis.

Chapter 4 provides a comprehensive analysis of the experimental results, focusing on the impact of tests, groups, and vowels on response accuracy across three perceptual tasks (identification, auditory discrimination, and category discrimination). Although production data was collected, it was not analysed due to time constraints.

Chapter 5 gives an in-depth discussion of the significant findings derived from the results chapter. These are discussed in the light of existing literature. Additionally, it identifies potential areas for future investigation, provides recommendations, and acknowledges any limitations encountered during the study.

Chapter 6 serves as a comprehensive summary of the entire research, highlighting key findings and insights.

Chapter 2. Literature Review

This chapter provides a thorough examination of the existing literature that shaped the establishment of the research questions for this thesis. The structure is as follows:

Section 2.1 addresses the perception and production goals that L2 learners should strive for.

Section 2.2 presents an overview of perceptual training, with a specific emphasis on the initial applications of HV training. This is followed by a thorough examination of the use of HV training in vowel learning, as well as an exploration of how accents are incorporated into speech perception training.

Section 2.3 delves into the impact of exposure to L1 on perceptual abilities. It examines how this exposure shapes these abilities, transitioning them from a universal phonetic sensitivity to a more specialised and language-specific state.

Section 2.4 is dedicated to the theoretical frameworks of L2 speech acquisition including the Perceptual Assimilation Model – PAM (Best, 1995), the Speech Learning Model – SLM (Flege, 1995), and the Second Language Linguistic Perception Model – L2LP (Escudero, 2005).

Section 2.5 provides an overview of vowel sounds (including definition, features, and classification).

Section 2.6 delves into the qualitative and quantitative characteristics of Arabic vowels, with a particular emphasis on the Saudi Arabian (SA) dialect, as well as the vowels in standardised varieties of English (British, Australian, and American).

Section 2.7 explores the perception and production of English vowels by Saudi Arabian learners, providing valuable insights into the difficulties and ease they experience with different vowels.

2.1 Understanding the essence of L2 perception and production

In the context of examining L2 learners' production of English vowels, contemporary academic literature places great emphasis on comparing their production to L1 speakers, commonly those who speak SSBE (e.g., Algethami, 2023; Alshangiti, 2015). The comparison is frequently conducted through the implementation of acoustic analysis (often vowel's midpoint), aiming to examine native-like phonetic patterns. This fundamentally contradicts the prevailing academic consensus: attaining a production similar to that of L1 speakers would be impossible for the vast majority of learners and only possible for a small group of highly motivated and talented individuals (Levis, 2005; Ioup et al., 1994). In light of this, what should learners strive for? Although the present thesis focuses exclusively on SA learners' difficulty with vowel perception, it proposes that any learners' production should be target-like, that is, intelligible to English speakers, agreeing with Ur (1996) who emphasises the significance of developing an understandable production that is suited to the needs and goals of the learner. Supporting this perspective, this study calls for the use of acoustic analysis not to adhere to strict native-like accuracy but to focus on the degree of changes and variation in learners' productions. By measuring variations in vowel realisations and assessing convergence among learner groups, acoustic analysis can provide deep insights into learners' progression before and after training. This methodology aligns with the broader goal of facilitating L2 acquisition by establishing realistic and meaningful targets for phonetic improvements.

Another effective method for assessing intelligibility is the vowel identification task (e.g., Iverson et al., 2012), which evaluates auditory intelligibility in both quiet and noisy conditions. In this task, L1 listeners (often SSBE) are required to identify learners' pronunciations from a set of closely related options. For instance, listeners are asked to identify learners' pronunciation of the word 'boat' from a closed set of options: 'boot', 'bait', and 'boat'. To evaluate target-like production, this task can be adapted to include both L1 speakers and proficient L2 listeners. This broader approach ensures an unbiased representation and confirms that learners' productions are understandable to individuals from both listener groups. In fact, the differences in acoustic characteristics between L2 speech patterns and those of L1 speakers do not necessarily imply a lack of intelligibility in learners' production. For example, in a study by Alshangiti (2015), various training approaches were investigated, including perception-based, production-based, and combined perception and production training. Arabic learners of English were assessed before and after the training in different tasks. For perception, tasks included Identification and discrimination. In terms of production, acoustic analysis was conducted. Despite the researcher's conclusion that training primarily affected its respective domain (perception training improving perception and production training improving production), it is interesting to see that all participants, regardless of the type of training they received, were judged to be intelligible to SSBE listeners during a vowel intelligibility assessment. When evaluating L2 sound production, the emphasis should therefore be on achieving intelligible production rather than striving for accuracy comparable to that of L1 speakers. The next section delves into L2 phonetic training, emphasising perceptual training, as it is the central focus of this thesis.

2.2 L2 Perceptual training

It can be challenging for language learners to identify and discriminate between some L2 phonological contrasts or sounds. For instance, they may struggle to distinguish between and/or produce the vowels in words like "check" and "chick" or "each" and "itch," resulting in miscommunication and affecting the intelligibility of their speech. Often, these difficulties arise because learners interpret L2 sounds through the lens of their nearest L1 sounds, or because the specific contrasts in L2 do not exist in their L1 (Flege & Port, 1981; Yamada & Tohkura, 1992). However, learners' perceptual patterns might not be set in stone since there has been neurophysiological and behavioural evidence for adult perceptual system flexibility (Norris et al., 2003). Since L2 learners do not lose their perceptual sensitivities, researchers have adopted auditory training over the last 20 years to improve L2 learners' perception and production of L2 sounds. When the FL is taught in an academic setting, exposure to the target language is frequently restricted, especially when students share the same native tongue. In light of this, L2 phonetic training can be a valuable resource in improving the understanding and production of difficult L2 sounds or contrasts, thereby facilitating L2 acquisition (Logan & Pruitt, 1995). This section provides a summary of the methodologies, designs, and tasks commonly involved in perceptual training. It then delves into an in-depth examination of the early applications of perceptual training, with a specific emphasis on HV training. Following this, studies that implement HV training in vowel learning are reviewed.

2.2.1. An Overview of the methodology and design of perceptual training

A basic design of the auditory training consists of a series of training sessions, and pre- and post-tests. Participants typically engage in a set of training sessions focused on identification (e.g., Logan et al., 1991; Wang & Munro, 2004), discrimination (Strange & Dittmann, 1984), or a combination of the two (e.g., Shinohara & Iverson, 2018; Carlet & Cebrian, 2019). The Identification task (ID) requires participants to identify or classify the stimulus they hear by selecting the option corresponding to the category they believe it belongs to from two or more alternatives. The primary purpose of this activity is to assess the participants' ability to classify and label auditory stimuli, which necessitates the use of memory and cognitive processes to assign the sound to an established category. As Identification training requires participants to categorise phonemes, it can improve representations of L2 sounds at the phonological level (Logan et al., 1991; Sadakata & McQueen, 2013). The discrimination task (DIS), on the other hand, assesses perceptual acuity and attention to detail. The task focuses on participants' ability to notice and discriminate minor differences between two sounds, which can improve their auditory-phonetic perceptual processing. It thus measures the ability to differentiate between minor acoustic fluctuations and form comparative evaluations of auditory stimuli (Logan & Pruitt, 1995). There are three generally recognised types of DIS tasks: ABX discrimination, AX discrimination, and oddity discrimination. The ABX discrimination task consists of three stimuli (A, B, X), and participants must determine whether X corresponds to A or B. The AX discrimination task requires the presentation of two stimuli, labelled 'A' and 'X', in a sequential fashion. Participants need to determine whether the second stimulus ('X') is similar to or dissimilar to the initial stimulus ('A'). For the oddity discrimination task, participants are exposed to three or more stimuli and are tasked with identifying the one that differs in category affiliation from the others.

After each trial in any training task, participants receive immediate feedback indicating whether their answers are correct or incorrect. This process allows individuals to identify the essential features of the sounds unconsciously and gradually (Logan et al., 1991). The duration of the training can vary significantly in studies, depending on the pace set by the researchers. The training period may range from 1-2 weeks (Iverson & Evans, 2009; Iverson et al., 2012), 3 weeks (Logan et al., 1991), or even extend to several weeks, such as 10 weeks (Shinohara & Iverson, 2018; Carlet & Cebrian, 2019). Prior to beginning the training, participants are required to complete a pre-test to assess their initial proficiency in challenging L2 sounds or contrasts. This sets a baseline for their performance. Following the training phase, a post-test is administered. There are two different forms that this post-test can take:

- **Using the same stimuli as the pre-test:** In this approach, the post-test includes the same stimuli as the pre-test. The primary objective here is to measure any significant improvements in performance from the pre-test to the post-test. These improvements are considered to be the direct result of the training.

- **Introducing new stimuli**: Alternatively, the post-test may feature new stimuli that were not part of the pre-test (e.g., Shinohara & Iverson, 2018). The test in this scenario assesses how well the training effects generalise to new words and/or speakers, thus measuring their capacity to apply their learning in different contexts.

If the post-test is identical to the pre-test, researchers conduct generalisation tests with novel stimuli not previously seen during training or testing. These tests, conducted following the post-test, are designed to evaluate if the training's effectiveness applies to different speakers and/or contexts beyond those in which the sounds were initially learned. For example, Lively et al. (1993) carried out two types of generalisation tests: one with novel words from an unfamiliar talker, and another with novel words from a familiar talker. To evaluate the extent of knowledge retention, it is common practice to administer post-tests and/or generalisation tests three to six months after the completion of the training (e.g., Lively et al., 1994). During these testing phases, the perceptual tasks typically involve either identification, discrimination, or a combination of both, without providing any form of feedback. A production test may also be incorporated to determine whether gains in perceptual abilities affect production capacities. The proportion of correct and incorrect responses determines the accuracy of identification and discrimination (Rato, 2014). Production progress is frequently examined by utilising evaluations by L1 English speakers (Carlet & Cebrian, 2019) or acoustic data analysis (as seen in Shinohara & Iverson, 2018).

2.2.2 The Initial applications of perceptual training

Strange and Dittmann (1984) conducted an earlier training study that provided the spark for subsequent research. Their investigation focused on supporting Japanese English learners in distinguishing between the phonetic contrasts of /l/ and /r/. This distinction poses a difficulty for these learners, as Japanese phonology involves a singular liquid whose acoustic properties are comparable to those of both /r/ and /l/. In order to modify the Japanese learners' perception of these sounds, a psychophysical method developed by Carney et al. (1977)[4] was implemented. This method entailed employing a consistent first stimulus in an AX pair over a sequence of trials while altering the second stimulus (i.e., using a different X each time). Using synthesised speech, Strange and Dittmann trained 8 Japanese learners[5] to distinguish between "rock" and "lock" sounds throughout 14 to 18 training sessions. The study's primary goal was to determine if the focused discrimination training could help learners effectively apply their acquired skills in different test scenarios, including identification tasks, minimal pair tests, and oddity discrimination[6]. In addition, the study sought to determine if training with a restricted set of synthetic stimuli containing /r/ and /l/ sounds in initial positions would effectively assist learners in developing phonetic categories for /r/ and /l/ in various phonetic contexts. The findings exhibited an ongoing

[4] Carney et al. showed that participants could discern minor distinctions within sound categories along a Voice Onset Time (VOT) continuum after undergoing several training sessions.

[5] They were students at an intermediate level of English, studying at the University of Minnesota.

[6] During the minimal-pairs test, participants chose the correct word from each pair they heard by circling it on their response forms. In identification tests, they labelled each sound as "R" or "L." For the oddity discrimination tests, they marked which one—first, second, or third—in each set of three stimuli was different.

increase in performance among all participants throughout the training sessions. Following the training, 7 out of 8 subjects effectively applied what they had learned to more complex identification and oddity discrimination tasks. In addition, the identification and discrimination of an acoustically dissimilar synthetic series consisting of the words "rake" and "lake" showed improvement in five out of seven subjects. Nevertheless, the improvement did not extend to natural speech, and training in a particular context failed to generalise to other settings, revealing limitations in the training's applicability to new knowledge.

Logan et al. (1991) performed a thorough review and improvement of the research methodologies utilised by Strange and Dittmann (1984) in an attempt to improve Japanese learners' ability to perceive the critical cues for the /r/ and /l/ sounds and to evaluate the training's efficacy in facilitating the generalisation of new words and speakers. Logan and colleagues identified two primary factors that contributed to the insufficient progress observed in Strange and Dittmann (1984), where participants faced difficulties in distinguishing between /r/ and /l/ sounds in natural speech tasks:

- AX Training Task: Strange and Dittmann utilised a fixed-standard AX discrimination task, which, as Logan et al. point out, could have led learners to rely primarily on basic sensory information in speech. It has been argued that this training, which emphasises sensory memory, may not be practical for broader applications, such as identifying minimal pairs.

- Choice of Training Stimuli: It was determined that relying solely on the /r/ and /l/ sounds from a single phonetic context was inadequate in preparing learners to deal with these sounds in various contexts. According to Logan et al., while Japanese learners may not generally differentiate between /r/ and /l/ in multiple contexts, past investigations (e.g., Sheldon & Strange 1982) indicate that they respond differently to these sounds depending on the phonetic environment. As a result, the training described by Strange and Dittmann (1984), which was limited to a singular environment, might not possess sufficient generalisability to alternative phonetic contexts.

Logan et al. (1991) replicated the training protocols Strange and Dittmann (1984) developed with significant modifications. These modifications are summarised in the following list:

1. **Stimulus variability**

   A variety of phonetic environments[7] were included in the stimuli, spoken by five different speakers. This was based on Posner and Keele's (1968) findings, showing that high variability in training stimuli leads to better performance with novel stimuli. Logan et al. also followed Strange and Dittmann's (1984) suggestion to use diverse phonetic contexts, hypothesising that Japanese learners would benefit from exposure to multiple speakers to overcome distinctive acoustic cues.

2. **Forced-choice ID task**

   A two-alternative forced-choice ID task was used during both training and testing phases. It was suggested that this task, by focusing on phonetic memory codes rather than sensory memory, would help participants better classify stimuli. Unlike the AX discrimination task, which focuses on detecting subtle acoustic differences, the ID task encourages broader classification (Jamieson & Morosan, 1986). Logan et al. also proposed that incorporating the ID task throughout the training, post-test, and generalization phases would improve knowledge transfer.

3. **Natural speech tokens**

   It has been suggested that training with natural speech can offer learners helpful cues for phonetic categories, enabling them to adjust their abilities for different speech situations. The provision of necessary cues may be lacking in synthetic

[7] The training stimuli consisted of 68 minimal pairs that contrasted the /l/ and /r/ sounds in different phonetic contexts. These included 25 **initial consonant clusters** (e.g., pray vs. play), 12 **initial singletons** (like rice vs. lice), **5 intervocalic words** (such as arrive vs. alive), 15 final singleton positions (e.g., rear vs. real), and 11 **word-final consonant clusters** (e.g., hoard vs. hold).

speech training, which could lead to restricted learning and challenges when applying acquired learning to unfamiliar speech
situations.

4. **The addition of two generalisation tasks (TG1, and TG2)**

   This was done to determine whether the acquired knowledge was effectively transmitted to new words spoken by an unfamiliar speaker (TG1) and a familiar speaker (TG2).

With the implementation of updated techniques, Logan et al. recruited a group of six Japanese learners[9] who completed a series of fifteen training sessions (each lasting 40 minutes) in a controlled laboratory setting for three weeks. Stimuli were presented in each session from one of five talkers, with a three-time rotation. The training employed a two-alternative ID task where learners had to identify contrasting pairs of /r/ and /l/ on a screen (68 minimal pairs). Immediate feedback was provided to guide their learning: correct responses moved them forward, while incorrect ones prompted a repetition of the pair and highlighted the right answer. The effectiveness of the training was evaluated by comparing participants' performance on pre- and post-tests. Note that the pre-test and post-test stimuli were recorded by a speaker not used in the training phase or the generalisation testing phase. The words in these tests were the same as those used in Strange and Dittmann's (1984) study. In these tests, participants were presented with 16 minimal pairs contrasting /r/ and /l/, each shown twice. Their task was to identify the correct word from each pair listed in an answer booklet (i.e., ID task). Following the post-test, three subjects[11] undertook two generalisation tests (TG1, TG2) to assess the applicability to novel stimuli. This consisted of 96 new words spoken by a speaker who had not been previously heard (TG1) and 98 new words spoken by a speaker who was familiar from the training phase (TG2). Note that no feedback was provided during the entire testing phase.

[9] All were Indiana University students who had lived in the U.S. for periods ranging from 6 months to 3 years at the time of assessment.

[11] Due to subject attrition, only three participants were able to take part in the generalisation tests.

The accuracy of identification showed a notable improvement from week 1 to week 2, but there was no statistically significant improvement from week 3 to week 4. Moderate yet significant progress was reported in the participants' overall identification accuracy, with a 7.8% gain on average from pre-tests to post-tests. Notable progress was also seen in the perception of words within initial clusters and intervocalic contexts, while initial singleton and final singleton positions experienced slight improvements; however, it is worth noting that the latter (final singleton) was already near the ceiling. Learners' performance in TG2 was 4% higher than in TG1, indicating that they were more successful when confronted with the familiar speaker in TG2. Relative to the pre-test, TG1 increased by 1.4%, while TG2 increased by 5.6%. In contrast, TG1 and TG2 achieved lesser scores of 6.4% and 2.2%, respectively, in comparison to the results of the post-test.

The results mentioned above underscore the advantages of the training, highlighting enhanced performance not only in the post-test but also in generalisation tests relative to the performance levels before the training began. This training procedure exhibited greater efficacy in comparison to the approach pioneered by Strange and Dittmann (1984). Logan and his colleagues concluded that their training successfully shifted Japanese listeners' attention to the critical cues of /r/-/l/ sounds. They attributed the fundamental success of their training to several factors: (1) the incorporation of highly variable stimuli (including a variety of phonetic contexts and talkers); (2) the utilisation of natural speech; (3) the consistent application of the same ID task throughout both the testing and training stages; and (4) the provision of immediate feedback following each trial throughout the training procedure. These methodologies, frequently employed in L2 phonetic training research, later became known as high-variability phonetic training (HVPT), high-variability (HV) training as per Iverson et al. (2012), or high variability production training according to Thomson (2018). Although there are variations in the specific methodologies employed in HV training, most studies typically include the following: presenting listeners with a variety of natural stimuli (from multiple speakers in diverse phonetic contexts), requiring them to complete ID tasks, and providing feedback on the accuracy of their responses (e.g., Iverson et al., 2005; Logan et al., 1991; Pruitt et al., 2006).

Lively and colleagues (1993; 1994) replicated the HV training and testing procedures originally developed by Logan et al. (1991) to assist Japanese learners/listeners in differentiating the English /r/-/l/ contrast, with each study concentrating on specific goals. Lively et al. (1993) aimed to train Japanese learners with high and low variability training conditions. In the first experiment, stimulus tokens from five talkers were utilised to train six learners to differentiate between /r/ and /l/ in various phonetic contexts—intervocalic positions, initial singletons, and initial consonant clusters. In the second experiment, a different group of six learners was trained to identify /r/ and /l/ from stimuli presented by a single speaker but across the five multiple phonetic environments initially used in the study by Logan et al. Pre-test and post-test procedures, as well as two generalisation tests[12], were employed to evaluate the impacts of each training condition. The findings from experiment 1 indicated a moderate yet significant improvement, with an average increase of 5.6% in mean accuracy from pre-test to post-test. In addition, learners could transfer their learning to new words produced by both familiar and unfamiliar speakers. They achieved a level of performance in the generalisation tests consistent with their performance during the final week of training (week 3). In contrast, the findings from experiment 2 revealed that despite making improvements in particular phonetic contexts between the pre-tests and post-tests, learners struggled to apply their acquired knowledge when encountering new words spoken by the novel speaker. In addition, their ability to apply their knowledge to new words from the familiar speaker was only moderate, as evidenced by their generalisation performance not exceeding that observed performance during the first week of training.

These results highlight the limitations of training with a single speaker, especially when testing generalisation. On the flip side, the advantages of training with multiple talkers are readily apparent, emphasising the importance of talker variability in achieving success in generalisation tests. Consistent findings were observed in the

[12] TG1 includes novel words produced by a familiar talker while TG2 includes novel words produced by a new talker.

research conducted by Wong (2012), wherein training methods with low and high variability were employed to improve the perception of English vowels /e/ and /æ/ among Cantonese learners. Both groups exhibited improved perception of vowels after ten training sessions under each condition. Yet, the group with high variability demonstrated higher proficiency in acquiring unfamiliar words and speakers in comparison to the group with low variability. The claim that HV training boosts generalisation is further confirmed by the investigation conducted by Kartushina and Martin (2019), in which they compared the speech production effects of high and low-variability training approaches. Their study found that only participants in the HV condition demonstrated generalisation to new talker stimuli and more consistent speech production.

Based on these findings, it is reasonable to hypothesise that learners who demonstrate improvement under typical HV training conditions —characterised by exposure to diverse phonetic contexts and multiple talkers of the same accent— would derive even greater benefit from high accent training, which incorporates multiple accent varieties. The rationale behind this hypothesis is that variability has been shown to enhance learning in typical HV training. By introducing accent variability alongside phonetic variability, learners are exposed to a broader range of linguistic inputs. This increased variability further challenges and engages the cognitive and perceptual mechanisms involved in language processing, leading to more robust phonetic representations and improved generalisation across different speakers and accents. In essence, if exposure to the variability of speakers and speech sounds within a single accent enhances learning, then incorporating multiple accents should foster a richer and more adaptive linguistic environment, improving phonetic performance through increased variability in accents, speakers, and contexts.

To the best of my knowledge, there has been no prior investigation into the effectiveness of the HV in accent training approach, which represents the primary focus of the current thesis. Nevertheless, it is necessary to acknowledge that Wong (2014) incorporated multiple English varieties into the HVPT approach, which included SSBE, Canadian, and American. The study, though, did not explore the effect of accent variability on learners' perception. Its primary objective was to assess the efficacy of various HV training approaches, including methods based on perception, production, and a hybrid approach combining the two. All training conditions featured multiple accents, making it difficult to determine whether the presence of a single accent or multiple accents would have resulted in different results.

Considering the potential impact of the small sample size of six Japanese learners in the studies conducted by Logan et al. (1991) and Lively et al. (1993), Lively et al. (1994) were prompted to develop an HV training programme that involved a larger group of monolingual Japanese listeners to enhance their differentiation of the /r/-/l/ phoneme. These participants, living in Japan, had limited English experience, mainly restricted to basic grammar training[14]. The researchers were also interested in the long-term retention of the training. 19 participants took part in the training (experimental group), whereas 23 completed pre- and post-tests without any training (control group). To ensure fair and unbiased comparisons between the experimental and control groups, it is advisable to provide the control group with training, such as transcription instruction. The methodology employed for both training and testing was identical to that of Logan et al. Following the post-test, participants underwent a subsequent assessment to evaluate the efficacy of the training on novel stimuli. This included two generalisation tests: 1TG, which used novel words from a new talker, and 2TG, which involved novel words from a familiar talker. In order to assess the degree of retention of the training, the post-test and generalisation assessments were administered again three and six months after the training's completion.

The control group did not demonstrate any significant difference in performance between the pre-test and post-test. This should be expected since participants did not undergo any form of training. Conversely, the experimental group exhibited a substantial 12% increase in performance, rising from 65% in the pre-test to 77% in the post-test. The decline in accuracy from the post-test administered after the training period to the post-test administered three months later was just 2%. No statistically

[14] Recall the individuals who took part in the investigations carried out by Logan et al. (1991) and Lively et al. (1993) were L2 Japanese learners residing in the US. However, Lively et al. (1994) included monolingual Japanese listeners with restricted exposure to the L2.

significant reduction in accuracy was noted for the generalisation assessments administered after three months or following the training. After six months, the accuracy in the generalisation tests remained 4.5% higher than the levels observed in the pre-test. Moreover, during the generalisation evaluations carried out six months later, there was no discernible decline in performance. Interestingly, generalisation accuracy to new words relied on the speaker's voice. After the training period, the accuracy of responses to new items generated by the familiar speaker was considerably higher than that of responses to tokens generated by the unfamiliar speaker (82% vs. 77%, respectively). Three months later, the results showed a slight edge for the familiar speaker, and by six months, this advantage for the familiar speaker had increased substantially.

The identification accuracy (12% from pre- to post-test) observed by Lively et al. (1994) exceeded the gains reported by both Logan et al. (1991) and Lively et al. (1993), which were 7.8% and 5.6%, respectively, despite all three studies employing the same methodology involving 15 HV training technique feedback sessions over three weeks. The difference in sample sizes may have been a contributing factor; Lively et al. (1994) had a larger group of 19 participants than Logan et al. (1991) and Lively et al. (1994), who each had six. Additionally, there was diversity among the participant populations; Lively et al. (1994) concentrated on Japanese listeners with limited L2 experience, a group that might have had more potential for progress than the Japanese L2 learners residing in the US in the research conducted by Logan et al. (1991) and Lively et al. (1993).

From the previously reviewed research, it is evident that the initial work of the HV training programme demonstrated improvement in the perception of the difficult consonants /r/ and /l/ by causing Japanese learners to shift their attentional focus towards more relevant phonetic features, extending their learning to new stimuli, and maintaining their learning over time (Logan et al., 1991; Lively et al., 1993; Lively et al., 1994). Despite this, there is insufficient evidence to robustly validate the concept of cue reweighting as the primary learning mechanism in typical HV training (Iverson et al., 2005; Iverson & Evans, 2009). For example, Iverson et al. (2005) investigated the effect of auditory training on cue weightings and the efficacy of various training approaches. Subjects were divided into separate groups, each receiving a distinct type of training. One group received HV training with real words from different speakers. The remaining groups were trained using modified natural speech recordings that had been modified through signal processing using three techniques[15]. Although Japanese adults improved across all training conditions, they did not change their acoustic cues to align with the modified stimuli. Rather, they showed improved accuracy in identifying stimuli as the English /l/ sound, particularly as the stimuli shared similarities with the Japanese flap sound, like a short closure and transition. It appears that engaging in auditory training can improve the ability to use prior knowledge for categorising speech sounds, even if the specific cues used in categorisation remain unchanged.

Aside from the /r/ and /l/ sounds, HV training methods have been shown to improve perception of a variety of other consonantal distinctions, all of which use ID tasks with immediate feedback: Pruitt et al. (2006) and Fuhrmeister & Myers (2017) successfully trained American English speakers on the Hindi dental and retroflex stops (the /d̪/-/ɖ/ contrast), Sadakata and McQueen (2013) taught Dutch learners on the Japanese geminate consonant contrasts, while Kim and Hazan (2010) provided training to British English speakers on the Korean /t/-/t^h/ contrast. Furthermore, the scope of this training extends beyond consonants, as it has also been tested on vowels. Even though vowels are less categorical than consonants (Pisoni, 1973), the approach has been utilised in various studies to support learners in establishing or reinforcing L2 vowel categories (e.g., Lengeris, 2018; Nishi & Kewley-Port, 2007; Iverson & Evans, 2009; Iverson et al., 2012). Aligning with this thesis's focus on vowel learning, the following section will examine the technique's application in vowel learning.

[15] *All Enhancement*, which enhanced F3 contrast and lengthened closure duration; *Perceptual Fading*, which gradually reduced F3 enhancement as training advanced; and *Secondary Cue Variability*, which increased variability in F2 and duration lengths during training.

2.2.3 High variability training in vowel learning

Iverson and Evans (2009) used HV auditory training to improve German and Spanish learners' perception of SSBE vowels[16]. Their study explored whether learners with small and large L1 vowel inventories (5 vowels in Spanish, 18 vowels in German) learn L2 vowels differently to gain insight into how existing L1 categories interfere with the acquisition of new vowels. Using L2 speech models such as SLM, the study hypothesised that German learners might find learning English vowels more difficult due to their dense L1 vowel space, leaving little room for new vowel categories. Spanish learners, on the other hand, with a smaller L1 inventory, would have more capacity to learn new vowels. Despite this, the study revisited findings from their previous research (Iverson & Evans, 2007), showing evidence that individuals with both large (German, Norwegian) and small (Spanish, French) L1 vowel systems perceived the L2 English vowels in similar ways, utilising the same acoustic cues[17], despite large overall effects of the L1—lower scores for smaller L1 systems and higher for larger ones. That is, there was little evidence that the language groups perceived English vowels in fundamentally different ways. However, Iverson and Evans (2009) pointed out that their earlier research 2007 focused on comparing how listeners from different L1 backgrounds perceive and categorise English vowels. Still, it did not investigate individual learning processes, such as through training.

In the initial experiment of Iverson and Evans (2009), 26 learners (13 Germans and 13 Spanish) received English vowel perception training through a UCL programme that included HVPT elements (multiple speakers, multiple phonetic contexts). Over 1-2 weeks, they completed five 45-minute sessions, each involving 225 trials of vowel identification with feedback. Each session began and concluded with 70 trials where participants identified 14 vowels, repeated five times in a random

[16] The study evaluated students with lower-intermediate English skills as determined by the grammar section of the Oxford Placement Test1. It compared Spanish speakers in London, who frequently encounter English, with German speakers in Germany, who have limited exposure beyond the classroom and media, but both groups were comparable in their ability to perceive English.

[17] It was found, for example, that all groups used essential acoustic cues like F1/F2 target formant frequencies along with subtle cues like formant movement and duration. This was observed even though Spanish and French vowels do not differ in formant movement and duration, while German and Norwegian vowels do.

sequence. The middle 85 trials were adaptively chosen based on the subject's specific errors, targeting frequently misidentified vowels. This design guaranteed comprehensive exposure to all vowels at the beginning and end, while personalizing the central training to each individual's needs. Participants were given a stimulus word, such as ‘slit’, and asked whether it sounded like ‘slight’, ‘slit’, ‘sleet’, or ‘slate’. For easier perception, each option was paired with a well-known word that shared the same vowel sound (e.g., ‘slight’ - 'night,' ‘slit’ - 'sit,' ‘sleet’ - 'seed,' ‘slate’ - 'eight'). Correct responses were rewarded with a "Yes!" display and a cash register sound, followed by a replay of the stimulus word. Incorrect answers resulted in a "Wrong" display, descending tones, and the correct and incorrect options being repeated. The training stimuli used 14 SSBE vowels in minimal pairs across four clusters (sets): 1) /e/, /ɑː/, /æ/, /ʌ/; 2) /iː/, /ɪ/, /aɪ/, /eɪ/; 3) /ɒ/, /əʊ/, /ɔː/; 4) /uː/, /aʊ/, /ɜː/ and were produced by multiple SSBE speakers (3 female, 2 male). The selection of these sets, which drew from Iverson & Evans (2007), was shown to be a challenge for learners with diverse L1 vowel systems, including Spanish and German. The first three sets were especially difficult for manly listeners, while the final one included the remaining vowels.

Participants were tested before and after perceptual training using natural recordings of /bVt/ words obtained from two British English speakers (1 female, 1 male); none of these stimuli or speakers were used during the training phase. The pre/post-tests included the following the following tasks:

1) Vowel identification

For the vowel identification task, participants were required to listen to the /bVt/ words and provide their responses using a closed-set format, which included all 14 words as potential response options. Participants submitted their answers by clicking on a button that displayed both the stimulus word and a familiar English word with the same vowel.

2) L1 assimilation

In the L1 assimilation task, participants were tasked to determine which of their native vowels most closely resembled the English vowel in the /bVt/ words they heard. They were instructed to categorise these L2 vowels as if they were listening to a native (L1) English speaker attempt to pronounce their own native language.

3) Vowel-space mapping

Participants were shown /bVt/ words (e.g., ‘bot’), along with common English words (e.g., ‘hot’). They heard synthesised /bVt/ words incorporated into a natural carrier phrase ("Say _ again," spoken by a male British English speaker). They were then asked to rate the degree to which the heard /bVt/ word matched an excellent example of the written word on a continuous scale. Participants provided their ratings by clicking along a continuous bar displayed on a computer screen.

After completing five HV training sessions, it was evident that both groups made significant progress in their vowel identification abilities. The average improvement for Spanish students was 10%, while German students showed a higher improvement rate of 20%, indicating that learning rates differed between groups. This finding suggests that, in contrast to what researchers assumed based on L2 speech models like the SLM, German learners' denser vowel space may have aided their learning, which is consistent with their previous research (Iverson & Evans, 2007), which found that German speakers outperformed Spanish speakers in learning. Spanish learners eventually reached a level comparable to German learners after completing ten additional sessions of the same technique in a subsequent experiment (with identical training procedures and pre/post-tests as in experiment 1). The study indicated that Spanish speakers, although initially slower in learning English vowels due to the influence of their smaller native vowel system, possess the same fundamental ability to learn English vowels as their German counterparts after receiving more extensive training. In light of this, the current thesis will carry out a lengthy training (16 training sessions, including two sessions per week) to train SA English learners. SA's vowel inventory is relatively small, with only 8 vowels (see section 2.6.1.2 on SA vowels). As a result, it is anticipated that SA learners will need an extended period of training to show improvement in vowel perception.

Although the learners made progress in identifying vowels, there were no significant differences in how closely the learners' 'best' vowel examples aligned with the average ideal English vowel examples in terms of first and second formant (F1/F2) positions, formant movement, and duration when comparing their 'best' vowel examples before and after training. The study argued that the HV training technique mainly enhanced the perception skills of Spanish and German learners by applying their existing knowledge of first and second language categories in naturally diverse speech, thus making their current categorisation processes more automatic and efficient. Despite the learners' initial knowledge of different language categories, the HV training improved their adaptability in unpredictable phonetic conditions without changing their perceptions of L2 categories, such as cue weightings. This is consistent with the findings of Iverson et al. (2005), who noticed that HV training did not affect the cues used by Japanese learners to perceive English /r/ and /l/ sounds. Iverson and Evans thus hypothesised that while HV training improves L2 categorisation skills like identification, it does not lead to the more profound phonetic perception changes that may occur during extended immersion in a language environment, such as living abroad. It is possible to argue that exclusive reliance on the ID task in HV training may not promote the development of new categories in adult learners, a conclusion supported by studies from Iverson et al. (2005) and Iverson & Evans (2009). Subsequent studies (e.g., Lengeris & Hazan, 2010; Alshangiti & Evans, 2015) used the ID task and found improvements in vowel identification but not in category discrimination. Combining discrimination and identification tasks in training may result in different outcomes for adult learners (See section 2.2.3.1 for the few studies examining the effectiveness of both types of training).

Using a large vowel set (14 SSBE vowels) in Iverson and Evans's (2009) study was deliberate, based on prior research suggesting the benefits of extensive vowel training. Lambacher et al. (2005) and Nishi & Kewley-Port (2007) found that Japanese adults learning English monophthongs improved significantly (16 to 25%). On the other hand, training English adults on Japanese vowel-length contrasts showed limited improvement, with little ability to adapt their learning to new speakers and phonetic contexts (Tajima et al., 2008). Nishi and Kewley-Port (2007) argued in favour of a broader approach to vowel training, suggesting that training on a more comprehensive range of vowels can lead to overall improvement. They found that Japanese learners who were trained on nine English vowels showed improvement across all categories, whereas those who were trained only on the three most challenging vowels only showed advancement in those specific categories. Recent studies by Thomson (2018) and Uchihara et al. (2021) also found that training in simple two-way contrasts (e.g., 'sit' vs 'seat') can be less effective. This is because, at least with vowels, learners' confusion is often more complex and not simply a matter of binary alternatives. Training on a broader range of sounds (e.g., 'sit', 'seat', 'set', 'sat', 'suit', 'sought') is preferred. Accordingly, this thesis trained L2 learners on a substantial number of vowels (16 L2 vowels), considering it beneficial for learners.

Much of the HV training research in vowel learning has focused on perceptual training and its potential impact on production (e.g., Brekelmans, 2020; Iverson et al., 2012). However, there is some evidence that combined perception and production training leads to better results in both perception and production (Alshangiti, 2015; Wong, 2014). Wong (2014) found that combining HVPT (akin to the methods in Logan et al.1991) with production training aided Cantonese English learners in distinguishing the /ɪ/- /iː/ contrast. During the production training, these learners observed videos of an L1 English speaker producing the vowel contrast in 20 actual words and explaining how the vowels are produced. Following that, the learners were instructed to practice producing these vowels these vowels three times with the researcher, who was an L2 English speaker, providing corrective feedback. Visual aids displaying articulatory details, such as tongue position (frontness or backness), were given in each session. Participants were also encouraged to use mirrors to help them improve their vowel production. The results indicated that the combined perception and production approach yielded higher levels of support for learners' production (66% accuracy) compared to production training alone (29%) or HVPT alone (36%). Production training alone did not help learners' perception, whereas other training approaches (HVPT, combined techniques) showed comparable efficacy in enhancing learners' perception.

Another study conducted by Alshangiti (2015) also found that combining perception and production training was more beneficial for Saudi Arabic learners of English in an immersive setting (London) than using either perception or production training independently. HVPT was used for perceptual training, and animations from the computer-Assisted Learning for Vowels Interface (CALVin) were used for production training. The CALVin program included diagrams of the primary articulators used in vowel production, like the lips, tongue, and jaw, presented in a schematic mid-sagittal view. Arabic learners viewed these animations at regular speed, listening to an L1 English speaker pronounce vowels in isolation and CVC words. Learners can record their utterances, listen to them back, and compare their recordings to the L1 speaker's. The goal was to activate learners' sensory feedback mechanisms, allowing for quick speech production adjustments, as Baker and Trofimovich (2006) described. Given the findings of Alshangiti (2015) and Wong (2014) that combining perception and production training is more effective than either method alone, this study predicts that the holistic approach (HV perception and production training) will improve the perceptual skills of Arabic learners. Importantly, this study addresses the combined method as a potentially more successful preference, but it does not provide empirical testing to verify its impact on perception skills in comparison with other methods[18]. Furthermore, even though production data was collected during the training and testing phases, they were not analysed in this study due to time constraints.

In contrast to the findings of Alshangiti (2015) and Wong (2014), who found that combining production with perceptual training does not hinder perceptual learning, Baese-Berk and Samuel (2016) conducted three experiments and found that production can disrupt perceptual learning. Their study focused on native Spanish speakers learning the Basque fricative contrast between /s̪a/ and /ʃa/. In the first experiment, 30 participants with no prior exposure to Basque were divided into two groups: one focused solely on perception and the other on both perception and production. All participants underwent a pre-test, training period, and post-test. In the

[18] To assess the impact of a combined approach on vowel perception in learners, different approaches should be developed and compared, including the combined approach as well as separate production and perception training techniques. The study's purpose was not to evaluate this; rather, it is assumed that the combined approach is a guaranteed preferred method.

pre- and post-tests, they completed 72 ABX discrimination trials involving sounds from three different continua (/s̻a/–/ʃa/, /s̻a/–/ts̻a/, and /s̻a/–/s̺a/). In each trial, participants heard two different tokens followed by a third token and were asked to determine whether the third token matched the first or second. No feedback was provided during these tests. The training involved five blocks of 72 trials (a total of 390 trials) on the /s̻a/–/ʃa/ continuum, with feedback given after each response. During perception-focused training, participants made ABX judgments without repeating the sounds. In the combined perception and production training, participants repeated the final token before making their judgment. The entire procedure, including the pre-test, the training, and the post-test, was conducted over two days, with a maximum of 48 hours between sessions. The findings demonstrated that participants in the combined perception and production group exhibited disrupted perceptual learning compared to those in the perception-focused group.

The second experiment replicated the training and testing procedures from Experiment 1, this time with 30 Spanish speakers who had intermediate proficiency in Basque. Although prior experience with Basque slightly reduced the disruptive effect of production, the group that combined perception and production still showed less robust learning compared to the perception-focused group. The third experiment aimed to further explore the source of the disruption observed in the previous experiments, specifically whether it was caused by producing the target sounds or by engaging the production system more broadly. In Experiment 3, 20 participants with no prior exposure to Basque underwent combined perception and production training. The procedures were similar to those in the combined perception and production groups in Experiments 1 and 2, with one key difference: instead of repeating the target speech sound after each trial, participants were asked to name a random letter (e.g., L, N, M) that appeared on the screen following the ABX discrimination task. These letters did not include the critical sounds being learned, ensuring that the production task was unrelated to the speech sounds being studied. The production of these unrelated sounds (letters) resulted in less disruption than producing the target sounds, but still interfered with perceptual learning, suggesting that the act of engaging the production system in general disrupts the development of perceptual representations.

The methodological differences between Baese-Berk and Samuel's (2016) study and those conducted by Wong (2014) and Alshangiti (2015) suggest that the disruption observed by Baese-Berk and Samuel may be more a result of task design than an inherent issue with production training. In Baese-Berk and Samuel's study, participants were required to produce sounds immediately after each perceptual task within the same trial, likely increasing cognitive load and hindering their ability to consolidate perceptual learning. This simultaneous engagement in production and perception within a single trial may have overwhelmed participants, limiting their capacity to effectively process the perceptual information. In contrast, Wong and Alshangiti reported that combined perception and production training yielded results comparable to perception-focused training, indicating no hindrance to perceptual learning. They employed high-variability perception training, involving identification tasks with immediate feedback, without requiring participants to produce sounds during each identification trial. Their production training was conducted independently of the perceptual tasks, ensuring that participants were not required to produce sounds within the same trial during perception. This approach enabled participants to focus on each task without the cognitive strain associated with switching between perception and production within a single trial.

While HV training is commonly carried out in laboratory settings, there is evidence indicating its effectiveness for L2 learners in both immersive and non-immersive (Wang & Munro, 2004; Iverson et al., 2012; Alshangiti & Evans, 2015; Iverson et al., 2023). For example, Iverson et al. (2012) examined whether the HVPT technique could improve advanced L2 learners' perception of SSBE vowels, considering these learners are already exposed to daily natural variations in English. They apply the method to two groups of French speakers: 15 high-proficiency learners in London and 21 low-proficiency learners in France. It was hypothesised that experienced learners in London would benefit little from training since their exposure to a wide range of L2 phonetic experiences in natural settings exceeds what can be provided in controlled training sessions. Learners completed Iverson and Evans's (2009) UCL training program, with modifications for this study, including more speakers (8) and sessions (8). The training lasted 1 to 2 weeks, with 45-minute sessions

Participants were evaluated on three tasks before and after training: vowel identification, category discrimination, and production accuracy. In the vowel identification task, participants heard one of the stimuli /bVt/ words and responded with a closed-set response that included all 14 words as response options. The category discrimination task required them to identify the odd one out of three words, two of which were the same and one of which was different, each voiced by a different talker. The production task required them to record each stimulus word three times before and after training. Following that, four British English listeners were chosen to judge learners' production. Reporting the perception results, the study found that both groups improved in vowel identification accuracy: experienced learners enhanced from 60% to 77%, and inexperienced learners improved from 41% to 66%. This demonstrates the distinct advantages of HV training, which is helpful in both second and foreign-language contexts. A modest improvement was noted in the vowel discrimination task in both contexts: experienced learners improved from 76% before to 83% after training, while inexperienced learners improved from 75% to 77%.

In their study, Alshanqiti and Evans (2015) investigated whether implementing CALVin-based production training[19] enhanced the perception and production of SSBE and if the learning environment influences its effectiveness. They tested 16 learners in London, UK (an immersive setting) and 9 in Jeddah, Saudi Arabia (a non-immersive setting). All participants completed five training sessions and four pre/post-tests, including vowel identification, category discrimination, speech perception in noise, and English vowel production. The results showed that training improved the production of some vowels (‘bit’, ‘bet’, and ‘bought’) in both groups (in London and Jeddah). However, the group trained in Jeddah, appeared to benefit more from training in

[19] It should be noted that the production training stimuli were recorded by a monolingual male speaker of SSBE, therefore, it does not qualify as HV production training. In contrast, the testing stimuli were recorded by 10 SSBE speakers (5 male, 5 female). The testing stimuli incorporated new words and speakers, not presented during training, to assess generalisation.

general and improved more in the production task, vowel identification, and noise speech perception. They were also more motivated than the group trained in London.

In light of the results of Alshangiti & Evans (2015) and Iverson et al. (2012), the present thesis predicts that HV perception and production training would benefit low-proficiency Arabic-speaking learners studying English in Saudi Arabia, where the exposure to English is limited to classroom settings. It could be advantageous to incorporate approaches or ideas (such as HV training) from speech research, known to enhance the perception and/or production of challenging L2 sounds, into classroom settings to enhance EFL teaching. Indeed, Barriuso and Hayes-Harb (2018) encourage curriculum designers and language teachers to use the HV technique's proven efficacy from laboratory research, as it appears to help learners better produce and perceive difficult L2 sounds. This is due to a continuing lack of educator expertise in methods that effectively improve students' production and perception skills (Thomson, 2018). Even with conventional teaching, perception and production training are frequently overlooked (Al-Nasser, 2015; Al-Shaibani, 2023; Olson, 2014; Foote et al., 2011). While L2 instructions reported that they have the ability and self-assurance to teach production, the majority reported lacking training or access to it and desired additional training in production instruction. Considering these shortages, integrating techniques proven to enhance learners' perception and production from L2 speech research into classroom practices would be ideal.

A recent study by Iverson et al. (2023) explored the use of HV training in vowel learning. Their approach differed from conventional HV training techniques, as they introduced a gamified approach to improve the perception of SSBE vowels among a study group that included 18 Japanese adult learners and 24 Spanish-Catalan children. Note that the reviewed training and assessment procedures and results apply exclusively to Japanese adults (experiment 1). Participants in Experiment 2 (Spanish-Catalan children) trained and were tested in a slightly different way (e.g., they trained with both unmarked and marked cards (audio-only version vs symbols plus audio version). For a thorough examination of Experiment 2 and its results, see Iverson et al. (2023). Adult students were trained online but tested in a classroom setting in

Spain. The game is played with a grid of shuffled, face-down playing cards. Learners click on two cards to display these; if they are similar, they are discarded; otherwise, they are refaced. The purpose is to locate pairs of matching playing cards by memorising their places from the previous round, which demands substantial memory. Following the match, the word's orthography is displayed. The game resembles a category discrimination task in that words must be matched among talkers without the need to identify the words, but it is played as an engaging experience as opposed to a trial-by-trail examination. Adult participants were instructed to play 200 online games independently for 10 days. The number of games played is displayed at the top of the screen and is tracked by the software. The game's training stimuli included multiple talkers (2 British males and 2 British females) and phonetic contexts identical to those used in previous HVPT studies (Iverson & Evans, 2009; Iverson et al., 2012).

Ten equal-gender British speakers who were not heard during the training phase recorded the testing stimuli. To assess the performance of the participants, the pre- and post-tests were administered:

1- *Vowel Identification test*

Each trial consisted of the display of a single stimulus and three or four answer alternatives. Participants were instructed to click on the word they considered correct without the ability to hear it again, yet they were informed of their correctness. A random speaker was chosen for 84 trials, which contained two repetitions, three noise conditions (without noise, with babbling, with a single talker masker) and fourteen vowels.

2- *Category discrimination test*

Each trial consisted of playing three words spoken by three British speakers concurrently with three animated frogs. Two words were identical, but one was different. Participants were required to select the odd word and received feedback on whether their answers were correct or incorrect.

Even though the memory game card more closely resembled the category discrimination task than the identification task, the results of experiment 1 revealed that Japanese adult learners did much better on the identification task than on the category discrimination test. The identification task is considered new since it was not done during training. The linguistic proficiency of the adult participants in this study is less informative. One possible explanation for the improved identification performance is that the learners' English competence is high, as indicated by the significant number of English words identified. These results demonstrate the method's efficacy for FL learners, even online. Notably, the LingLab vowel matching game[21] established for this thesis draws inspiration from the memory-card vowel training used in this study (Iverson et al., 2023). Still, it further incorporates a variety of L1 and L2 English varieties.

2.2.3.1 High variability identification and discrimination training

Previous research on HV training, which included ID tasks, revealed limited or no improvements in vowel discrimination accuracy following the training despite improvements in vowel identification accuracy (Lengeris & Hazan, 2010; Iverson et al., 2012; Alshangiti & Evans, 2015). A handful of studies have been undertaken to assess the effectiveness of DIS and ID tasks in HV training (Carlet & Cebrian, 2019; Shinohara & Iverson, 2018; Nozawa, 2015; Wayland & Li, 2008; Flege, 1995).

In their study, Carlet and Cebrian (2019) researched HV training, explicitly exploring the efficacy of DIS and ID tasks. The objective was to enhance the perception and production of five SSBE vowels /iː ɪ æ ʌ ɜː/, which have been determined as difficult for Spanish/Catalan learners of English. 54 learners underwent five 30-minute sessions over 10 weeks. They were divided into three groups: 18 trained in the AX discrimination task (nonsense words) with feedback, selecting either "same" or "different", 20 in a seven-option forced-choice identification task (nonsense words) with feedback, and 16 in the control group engaged in transcription practice using a web tool. The vowel identification tests, featuring both actual and nonsense words spoken by new speakers, were carried out before and after the training. The purpose of these tests was twofold: the nonsense words were used to evaluate the

[21] The game was designed with the assistance of the Research Software Engineering (RSE) Unit at Newcastle and under the guidance of Dr Ghada Khattab

impact of the training, while the real words were analysed to assess generalisation to actual language use. A week after the post-test, a follow-up generalisation test was administered, which included new non-words spoken by familiar speakers that participants had encountered during the training. In addition, a delayed post-test (i.e., a retention test) was conducted two months after the initial post-test to evaluate the lasting effects of the training. Correct identification scores were calculated for each testing phase for analysis. A picture naming task was conducted both before and after training to evaluate L2 production[22].

The production results showed no progress for the control group from the pre-test to the post-test. Nevertheless, the training groups experienced an improvement in production ratings after the training, with the DIS group's median scores increasing by 0.4 and the ID group by 0.6. Similarly, the perceptual results showed that the two training groups (ID, AX-DIS) contribute significantly to L2 learning. When comparing pre- to post-test vowel identification accuracy (i.e., non-words stimuli), the ID group improved by 26.3%, while the categorical DIS group improved by 9.8%. Both experimental groups outperformed the control group in terms of their ability to generalise to new nonsense words. There was no significant difference in the ID group's post-test and generalisation scores, whereas the DIS group's generalisation scores were significantly higher than their post-test results. The ID group, on the other hand, demonstrated more effective generalisation to the perception[23] and production of real-world stimuli. The delayed post-test results were consistent with the post-test results, indicating that both groups retained their learning two months after training completion.

The researcher suggested the more remarkable improvement in the ID group's performance was due to task familiarity, as they were tested with identification tasks before and after training. The study's limitation includes the absence of the DIS test, which may disadvantage the DIS group in task familiarity compared to the ID group.

[22] Participants identified 27 pictures (real words) and repeated each word twice. The production was evaluated through the use of judgments from four SSBE speakers: they were asked to identify the sound they heard and rate it on a 9-point Likert scale.

[23] The ID group improved by 15%, while the DIS group improved by only 1.5%.

However, Flege (1995) and Carlet (2017), who only used identification tasks to assess DIS and ID training effects, found no significant difference between DIS and ID groups in perceiving English stop consonants, indicating no task familiarity effect. Lacking the DIS test in Carlet and Cebrian (2019) can hinder straightforward comparisons to studies employing ID tasks during training and assessing learners with both ID and DIS tasks before and after training (e.g., Alshangiti & Evans, 2015; Iverson et al., 2012). Yet, since discrimination training improves vowel identification, as shown by Carlet and Cebrian (2019), it is reasonable to expect improvements in vowel discrimination as well.

Shinohara and Iverson (2018) assessed the effectiveness of DIS training by comparing conventional ID training to DIS training. They also investigated whether using different focus methods (categorisation versus noticing subtle differences) could lead to improved learning outcomes. It was hypothesised that HV discrimination training could focus on perceptual aspects more than identification training. Japanese speakers' mistuned auditory-phonetic perceptual processing could have a role in their challenges with learning the /r/–/l/ sounds. Conversely, HV identification training could improve learners' phonological representations of English /r/–/l/ by having them categorise these phonemes. 41 Japanese students took part in a study designed to enhance their perception and production of the English /r-l/ contrast. They underwent 10 training sessions over a period of 10 to 28 days, hearing a different speaker in each session. The sessions were equally divided between ID and DIS tasks, with 5 sessions allocated to each. Half of the participants started with discrimination training (DIS - ID), while the other half began with identification training (ID - DIS). Following each training session, participants took a short identification test using stimuli from the same speaker who had spoken during that session. Their percentage of correct answers was displayed at the end of the training to track their progress. In the ID training, participants undertook a two-choice task, selecting one option from a minimal pair (like 'rock-lock') based on a single auditory example, and received instant feedback for each selection.

Participants in the DIS training completed three tasks: natural and signal-processed auditory discrimination (AD), as well as category discrimination (CD), all with a three-option forced choice and immediate feedback. The first AD task required the identification of the odd word (e.g., *'red, red, led'*) using natural minimal pair recordings (the same speaker produced the three words). The second AD ask involved signal-processed stimuli with distinct F3 variations to identify F3 differences. The final task, CD, presented three stimuli where participants discerned the one with a different initial phoneme, ignoring other acoustic variations (e.g., choosing 'red' from *'red, link, light'*). Participants were evaluated three times to assess training effects: at the start (pre-test), after the first five sessions (mid-test), and after the second set of five sessions (post-test), with no feedback provided during these tests. The perception tests consisted of three tasks: ID, AD, and CD[24]. Identification and discrimination were assessed through correct/incorrect responses. Participants completed two tasks for production: a word-reading task involving 40 minimal-pair words and a passage-reading task involving an excerpt from "The Rainbow Passage" (Fairbanks, 1960). Praat software was used to objectively measure the F3 frequencies of target tokens.

The study showed that both HV identification and discrimination training were practical, resulting in comparable improvements. Both methods boosted learners' identification, auditory and category discrimination skills, and production of the /l/-/r/ contrast, with identification training showing a slight edge in the post-test. The study concluded that as long as ID and DIS training methods include high variability elements, the effects on L2 learners remain comparable. Moreover, no significant differences were found between the training methods in the brief identification tests at the end of each session. The improvement in identification observed in this study was around 12%, which was not greater than the improvement seen in standard HV training that only used ID tasks (e.g., Iverson et al., 2005; Lively et al., 1993). Based on these findings, researchers concluded that identification and discrimination training have similar effects on fundamental mechanisms, though the specific nature of the improved processes is undetermined. They provided three possibilities:

[24] The ID and CD tests followed the same procedures as the training tasks. The AD test consisted of three synthetic stimuli focusing on F3 and F2 sensitivities at the boundary, as well as F3 within the /r/ category, with participants identifying the distinct stimulus.

1- One possibility was that both ID and DIS approaches enhanced auditory-phonetic awareness, particularly by increasing sensitivity to F3 at the /r-l/ boundary and improving accuracy in category discrimination. This improvement, however, was not particularly discerning, as there was also a rise in F2 sensitivity at the phoneme boundary, which is considered a less significant cue.

2- Another possible explanation was that the participants' internal process of classifying phonemes, or phonological labelling, could have influenced their performance in the category discrimination task. This conclusion was supported by the observed parallel between participants' improved ability to discriminate between sound categories and their increased accuracy in labelling phonemes.

3- The last possibility was that both training methods may have aided listeners in dealing with a variety of stimuli rather than learning new categories of English /r/ and /l/ sounds. This is consistent with the explanation proposed in previous research (e.g., Iverson et al., 2005), which found that HV identification training improved Japanese speakers' perception of English /r/ and /l/ but did not change their basic category representations. They became more automatic and consistent at applying their L1 flap /r/ category to English. This implies that the training reinforced pre-existing sound categories rather than adjusting them to align with new training cues.

Shinohara and Iverson (2021) conducted a training study that closely paralleled their earlier research from 2018, utilising similar training procedures. However, a notable distinction in the 2021 study was the exclusive use of synthesised stimuli in the AD task, featuring signal-processed variations in the F3 formant, in contrast to the 2018 study which included two AD tasks—one with natural words and another with synthesised stimuli. The primary focus of their research was to explore how age affects Japanese learners' ability to discriminate the English contrast /r/ and /l/. While the study was not primarily intended to evaluate the overall effectiveness of the identification and discrimination tasks, the findings strongly reinforced the efficacy of these training methods. The training group included 47 participants, ranging from children to adults. The study anticipated that children would benefit the most from the training. However, the results showed improvements in both perception and production across all age groups, with adolescents achieving the most significant gains, particularly in identification, category discrimination, and sensitivity to the F3 formant. This challenges the "younger is better" hypothesis (Flege, 1995), as younger children did not outperform adolescents. While age influences the extent of improvement, the effectiveness of HV identification and discrimination training is evident across different age groups.

It can be concluded that both HV discrimination and identification training are effective, enhancing not only identification but also auditory and category discrimination skills. However, the exact process being improved is not fully understood, as Shinohara and Iverson (2018) pointed out. This finding differs from previous research, which found that HV identification training alone improved identification but did not improve discrimination (e.g., Heeren & Schouten, 2008; Iverson & Evans, 2009; Lengeris & Hazan, 2010; Alshangiti, 2015). It is possible that, at least for vowel learning, HV discrimination is needed to see improvement in vowel discrimination. The current thesis intends to use a combination of perceptual tasks (ID, AD, CD) in both the training and testing phases, thus encompassing both HV identification and discrimination training, with the expectation that both will benefit learners. Moreover, given the three-month duration of these phases, including a variety of activities is intended to improve the learning experience and keep participants engaged throughout the extended training period.

2.2.4 Variability of accent in speech perception training

In L2 training research, the effect of accent variability (exposure to various accents) on speech perception has received insufficient consideration. Nonetheless, a limited number of L1 training studies have investigated the impact of accent variability on speech perception. Baese-Berk et al. (2013) examined how monolingual American English speakers perceive speech with foreign accents from different linguistic backgrounds (Thai, Korean, Hindi, Romanian, and Mandarin) when undergoing systematic variability training. It was believed that by being exposed to various foreign accents, L1 English speakers could acquire the ability to identify shared patterns (commonalities), such as a slower speaking rate. These commonalities[25] provide a predictable variation that helps listeners understand unfamiliar accents. Hence, the researchers hypothesised that if listeners could adjust to the variability of individual speakers with multiple foreign accents (one speaker per accent), they would similarly adapt to a novel accent and a novel speaker with a previously encountered accent.

Participants were given two training sessions over two days. During each training session, participants were exposed to 160 sentences in five repetitions of a Bamford–Kowal–Bench (BKB) sentence list (Bench & Bamford, 1979). Each repetition featured sentences spoken by an individual speaker with a distinct linguistic background (i.e., Thai, Korean, Hindi, Romanian, and Mandarin) for five speakers. The participants were instructed to transcribe the sentences they heard, and no feedback was provided to them during this process. The post-test was administered right after the second training session. In the post-test, participants were presented with a set of BKB sentences mixed with white noise (+5 dB signal-to-noise ratio) from a new Mandarin-accented English speaker (familiar accent) and a set of BKB sentences from a new Slovakian-accented English speaker (unfamiliar accent). In the same manner as the training, they were instructed to transcribe the sentences they heard without feedback.

Baese-Berk et al. (2013) compared the performance outcomes of their study's group, which was trained with a variety of foreign accents, with two groups from Bradlow and Bent (2008): one trained exclusively in American English (without any foreign accent) and the other trained in Chinese-accented English. This comparison was enabled by consistent training methods, post-test procedures, and the use of the same speakers and sentences in the post-tests across all groups. A direct comparison was possible due to the consistent training methods, post-test procedures, and the same speakers and sentences in the post-tests across all the groups.

[25] Yet, the researchers did not report which specific commonalities native English speakers have learned through training that incorporates various foreign dialects.

The findings of Baese-Berk et al. (2013) demonstrated that training with multiple foreign accents enables generalisation to new speakers with the same trained accent (Mandarian) as well as to a new speaker with a novel accent (Slovakian). This results in a generalisation that is both speaker and accent independent, highlighting the benefits of exposure to a variety of accents in improving phonetic adaptability and performance. In contrast, the performance of the control and single accent groups from Bradlow and Bent (2008) was significantly lower than the multi-accent group when exposed to the novel Slovakian accent. The single accent group performed similarly to the multi-accent group and higher than the control group when hearing a new speaker of a familiar accent (Mandarin), indicating a generalisation that was independent of the speaker but dependent on the accent.

The findings of the studies reviewed in this section (Baese-Berk et al. 2013, Bradlow & Bent 2008), demonstrate that exposing L1 listeners to a variety of foreign accents leads to better generalisation to an unfamiliar accent (Slovakian). This benefit was not observed when exposure was limited to a single accent. Based on these observations, L2 learners might experience similar benefits from exposure to a range of English accents. Although the L1 listeners in the Baese-Berk et al. were trained in their native language, differing from L2 learners who are exposed to a second language, the exposure to various English accents would provide essential acoustic cues. Such exposure might enhance learners’ perception of various speech patterns, better equipping them to handle encounters with unfamiliar accents. Therefore, it would be interesting to explore whether L2 learners can gain any positive benefits from incorporating various accents into the HV training technique.

The following list provides a concise summary of key findings from the training studies covered in section 2.2, forming the basis for the hypotheses of this thesis.

1. **The Role of Accent Training**

- Finding: Previous research (Logan et al., 1991; Bradlow & Bent, 2003) indicates that HV training improves learners' perception and production.

- *Hypothesis*: HV accent training will also improve learning outcomes, with students exposed to multiple accents outperforming those who hear only one accent.

2. EFL Instruction

- Finding: Iverson et al. (2012) demonstrated that HV perception training improved French learners' ability to perceive English vowels, while Alshangiti and Evans (2015) found that CALVin-based production training helped Saudi learners improve their pronunciation of English vowels. Both studies were conducted in non-immersive environments (France and Saudi Arabia).
- *Hypothesis*: HV accent training is expected to improve EFL teaching by helping Arabic-speaking learners in non-immersive settings better identify and distinguish challenging L2 vowels.

3. Training Focus

- Finding: Alshangiti (2015) and Wong (2014) found that combining perception and production training is more effective than using either method alone for improving L2 learners' skills.
- *Hypothesis*: Arabic learners are expected to benefit from this combined approach, though this thesis does not test it directly, but suggests it as the more effective method based on existing literature.

4. Perceptual Tasks

- Finding**:** Carlet and Cebrian (2019) and Shinohara and Iverson (2018) demonstrated that both HV identification and discrimination training improved learners' vowel identification and discrimination abilities
- *Hypothesis*: HV identification and discrimination training will improve Arabic-speaking learners' ability to perceive English vowels.

5. Number of Training Sessions

- Finding: Iverson and Evans (2009) observed that Spanish learners' vowel perception improved moderately after five HV sessions, while German learners showed a 20% improvement. After 15 sessions, both groups performed similarly, suggesting that learners with simpler vowel systems (Spanish) may need more training than those with more complex systems (German).

- *Hypothesis*: Arabic learners, with a simpler vowel system, are likely to require extensive training to effectively acquire English vowels.

6. **Target-Trained Vowels**

- Thomson (2018) and Uchihara et al. (2021) found that training with two-way contrasts (e.g., 'sit' vs. 'seat') is ineffective for vowels, as learner confusion patterns are rarely binary. Training on a broader range of sounds is more effective.
- *Hypothesis*: Arabic learners will benefit from exposure to a broader range of vowel sounds.

This section reviews studies on L2 phonetic training, explicitly highlighting the use of the HV training method in vowel learning. The following section examines how prior exposure to a native language can shape one's perceptual abilities, shifting from a general to a more language-specific perspective.

2.3 Language-specific speech perception and production

In the very beginning stages of **infancy**, infants demonstrate the ability to distinguish phonetic properties that differentiate phonetic segments across languages. This universal phonetic sensitivity begins to diminish around six months of age when exposure to their native language begins to influence their perceptual abilities and make them more language specific. This transition occurs more quickly for vowels than consonants. By the end of their first year, infants' perceptual abilities more closely resemble those of adults, showing the same limitations in distinguishing the L2 phonetic sounds or contrasts. As individuals become more adept at perceiving sounds from their native language, their ability to distinguish sounds from other languages declines. This proficiency in native sounds (becoming highly sensitive to specific native speech cues) influences how they acquire the L2 sound later in life, mainly when one or both members of an L2 contrast are realised differently than in the learner's L1 (Werker & Tees, 1984).

Numerous studies have examined the shift from a language-general to a language-specific perspective. For instance, Bosch and Sebastián-Gallés (2003) examined 4- and 8-month-old infants from monolingual Spanish, monolingual Catalan, and bilingual Spanish-Catalan backgrounds. They tested their ability to distinguish Catalan vowel contrast (/e/-/ɛ/). It was observed that infants of a younger age could perceive this vowel distinction regardless of their linguistic background. Several studies, including Best (1995) and (Werker & Curtin, 2005), have confirmed this pattern, in which younger infants can perceive contrasts that are not present in their native language, but older infants have difficulty. Theories on L1 speech perception postulate that exposure to one's L1 causes a shift from a general to a language-specific perception of the language (Werker & Curtin, 2005). The Native Language Magnet theory (NLM) (Kuhl, 1994; Kuhl et al., 2008) supports this proposal. According to Kuhl (2004), the more infants are exposed to their L1, the more dedicated their brains become to that language. This perceptual shift, known as the perceptual magnet effect or "perceptual warping", diminishes their ability to perceive sounds of non-native languages while enhancing the processing of their native language (Refer to section 2.4.1, where the theory was developed to explain the perceptual patterns of L2 adult learners).

While infants acquire their L1 with relative ease, **adult learners** often struggle with L2 acquisition. This difficulty was traditionally attributed to the Critical Period Hypothesis (CPH), which suggested that brain changes after puberty impede a person's ability to acquire their L2 in the same manner as the L1 (Lenneberg, 1967; Munro et al., 1996). Recent research on L2 acquisition, however, has produced inconsistent findings regarding the existence of a critical period for language acquisition. For example, Flege and colleagues (1995) looked at the English proficiency of native Italian speakers who had lived in Canada for around 32 years. These participants were asked to record five brief sentences, which were then assessed for foreign accents by L1 English speakers. Contrary to the CPH, the results of this study demonstrated that as the participants' age of arrival (AOA) in Canada increased, the perceived strength of their accents decreased. These findings suggest a linear correlation between AOA and the strength of one's accent. It is possible that a biological "cut-off" age does not cause a decline in language acquisition with age. Instead, it may be that the L1 learned early in life influences and hinders the ability to learn additional languages later in life. The more familiar a person is with their L1 language, the more challenging it is to learn and distinguish L2 sounds. In light of this, it is clear that L1 should have a significant impact on learners' perception and production of L2 sounds. The upcoming section reviews the most influential L2 speech theoretical models (Kuhl, 2000; Best, 1995; Flege, 1995; Escudero, 2005) that mainly link the difficulty of perceiving and producing L2 sounds to the similarity between L1 and L2 sounds. The results of the current thesis pertaining to vowels (whether they are simple or challenging) will be discussed in relation to these L2 speech frameworks' predictions.

2.4 Overview of second language Acquisition theoretical frameworks

The most influential theories of L2 speech perception are the Native Language Magnet theory –NLM (Kuhl (2000), Perceptual Assimilation Model – PAM (Best, 1995), Speech Learning Model – SLM (Flege, 1995), the Revised Speech Learning Model – SLM-r (Flege & Bohn, 2021), and Second Language Linguistic Perception Model – L2LP (Escudero, 2005). While the SLM-r suggests that L2 segmental production and perception evolve simultaneously without one necessarily preceding the other, the other models attribute incorrect production to inaccurate representation of L2 sounds. In other words, L2 production errors are predominantly perceptually driven. Following these predictions, much effort has been made to improve the perception of difficult L2 contrasts/ sounds, that is, the ones that assimilate to a single native category (e.g., Lambacher et al., 2005; Lively et al., 1994). This may explain why the majority of training studies are perceptual rather than production–based. For example, (Flege, 2003) claimed, "L2 phonetic segments cannot be produced accurately unless they are perceived accurately" (p.27). If learners do not perceive the phonemic categories of the L2 correctly, they cannot produce them correctly.

However, this concept that perception precedes production has been tested with conflicting results on various L2 learners. For example, Sheldon and Strange (1982) found that Japanese-speaking learners accurately produce the English segments /l/-/r/ despite their poor perception. In contrast, Ingram and Park (1997) found that Korean and Japanese learners could not correctly produce the English vowels /e/ and /æ/ due to their inaccurate perception. These findings indicate that improving learners' perception and production are interrelated and crucial components of the learning process. Additionally, L2 speech models acknowledge the possibility for learners to attain native-like L2 categories when they receive enough exposure to the target language (e.g., Flege, 1995). The level to which a learner can accomplish target-like categories, as opposed to native-like categories, is an underexplored area of research that requires deeper investigation. In addition, it is crucial to establish the criteria for defining target-like perception and production. While it is generally agreed upon that the similarity between L1 and L2 sounds influences second language acquisition, the specifics, such as which L2 sounds (new, similar) learners find challenging or straightforward, are inconsistent and subject to variation based on the assumptions made by different models. The following section summarises the postulates and assumptions of L2 speech models (NLM, PAM, SLM, SLM-r, L2LP)[26].

2.4.1 Native Language Magnet Theory (NLM)

As highlighted in section 2.3, the NLM theory was introduced by Kuhl in 1994 to comprehend speech perception from a cognitive standpoint. It posits that during their first year, infants construct 'prototypes' or standard representations of sound categories from their native language. These prototypes serve as a reference point for other sounds within the same category. The NLM theory expands its scope to include how adults perceive sounds when acquiring a second or foreign language. According to Kuhl (2000), the distinctive auditory characteristics of learners' L1 influence how they interpret L2 sounds. Influenced by the prototypes of their native language, L2 learners may find it difficult to differentiate between sounds closely related to those of

[26] These models were chosen due to their recognised relevance in L2 acquisition research. The thesis examines the findings on easily distinguishable and challenging, taking into account their phonological and phonetic similarities and differences.

their L1 (similar sounds). On the other hand, it would presumably be easier for them to distinguish sounds that differ significantly from their native prototype.

For example, the NLM theory explains why Japanese students learning English frequently struggle to differentiate between the English sounds /ɹ/ and /l/. This phenomenon is attributed to the interference of Japanese linguistic prototypes with English sound categories, causing the perception and production of these sounds to become confused (e.g., Kuhl & Iverson, 1995). In their study, Iverson et al. (2003) investigated how Japanese, German, and American English speakers perceived synthesised /r/ and /l/ sounds whose F2 and F3 properties were modified. Participants were asked to classify these sounds based on their L1 phonemes and then evaluate how closely these categorisations aligned with their L1 sounds. During a discrimination task, they also distinguished between sound pairs that differed solely in their F3 property, with half being similar and the other half being dissimilar. The study revealed that Japanese participants, who typically categorise both /r/ and /l/ as one, had difficulty distinguishing F3 differences but were more responsive to F2 differences, which is not essential for /r/ and /l/ distinction. Iverson and colleagues believe that these Japanese listeners, based on their perceptual understanding, leaned on the less relevant F2 cue to interpret /r/ and /l/ sounds. This supports the NLM theory that the perception of L2 phonetic categories is shaped by native language prototypes.

Although the NLM theory is not primarily a model of L2 learning, it proposes that increased exposure to the L1 strengthens neural commitment as L1 is consistently reinforced throughout an individual's lifespan. On the other hand, increased exposure to a L2 can modify perceptual representations. Therefore, NLM provides valuable insights into the flexibility of language learning and the ways individuals adapt to new linguistic environments.

2.4.2 Perceptual Assimilation Model – PAM & PAM-L2

As Best (1995) proposed, PAM is founded on the Direct Realist view and Ecological Theory of Perception (Fowler, 1986). Additionally, it draws upon the Articulatory Phonology framework developed by Browman and Goldstein in 1989. According to the

Direct Realist perspective, individuals acquire temporal and spatial information from the environment through integrated perceptual systems without relying on inherent knowledge or acquired mental associations. It is suggested that listeners perceive information about the position and motion of speech articulators (like the tongue and lips) during sound production without involving cognitive mechanisms or mental representations because the environment provides abundant and direct information. In the context of L1 acquisition, infants decode information from the articulatory movements that lack linguistic content. Utilising integrated perceptual systems such as hearing and vision, they detect high-level constant features—articulatory information that conveys meaning in their native language. This detection mechanism becomes more refined throughout life, becoming more accurate at discriminating critical distinctions in native speech patterns.

Adult naïve listeners

PAM is primarily a perceptual model that predicts the assimilation patterns of L2 contrasts and thus does not provide predictions regarding the production patterns. It describes how adult naïve or early L2 listeners assimilate the contrasts of non-native speech sounds and anticipate their capacity for accurate discrimination based on this assimilation. Adhering to the Direct Realist view, PAM claims that adults perceive articulatory actions (such as tongue and lip movements) of speakers directly, without employing cognitive representations for auditory perception, and identify invariants from them. The model proposes that listeners classify non-native sounds based on their similarity or dissimilarity to the sounds in their native language, using their innate phonological system to perceive and classify unfamiliar non-native language sounds. During this auditory process, non-native listeners associate novel sounds from the target language with sounds from their native language that are most similar. Thus, the PAM evaluates the similarity between L1 and L2 sounds based on their articulatory and gestural properties. The following discussion presents six distinct categories of assimilation types anticipated by the PAM for non-native listeners when they encounter unfamiliar language sounds. These categories show varying degrees of perceptual difficulties seen in L2 phoneme discrimination (Best, 1993). Each category proposes a prediction about a listener's perceived difficulty or simplicity in discerning L2 sounds based on how they are mapped into the L1 system.

Firstly, a **"two-category (TC) contrast"** occurs when two L2 sounds correspond to two different L1 sounds, making it easy to discriminate between the two L2 phonemes. For instance, Spanish learners of English perceive the English sounds /i/ and /u/ as two separate native sounds (Escudero & Boersma, 2004). The assimilation of Japanese /w/ and /j/ to their American English counterparts /w/ and /j/ is another example. A **"category-goodness (CG) difference"** occurs when two L2 sounds are assimilated to a single L1 sound; however, one of the L2 sounds is a better example of the L1 sound than the other. In this case, discrimination would be very good to moderate. For example, Greek-speaking learners tend to assimilate the English sounds /ʌ/ and /æ/ to the corresponding Greek /a/ sound. They perceive one of these English sounds as having a more remarkable resemblance to the Greek /a/ than the other (Lengeris, 2009). Another proposed type of contrast is referred to as "**single-category (SC) assimilation"**. In this case, two L2 sounds are assimilated into a single category in the L1, where both phonemes are considered to be good examples of the L1 sound (e.g., the English sounds /i/ and /ɪ/ are perceived as Arabic /i/). This would create a considerable challenge in discerning the L2 sounds (weak discrimination).

On the other hand, the "**both-uncategorised (UU) type"** is suggested when the listener cannot find any similar sound for the two L2 phonemes; as a result, discrimination is predicted to range from poor to moderate. For example, Japanese listeners do not differentiate between English nasal sounds /ŋ/ and /n/ when they appear at the end of a syllable, resulting in both being unclassified (Aoyama, 2003). In the "**uncategorised-categorised (UC) pattern**", it is anticipated that the discrimination will be excellent in a scenario where only one of the two L2 sounds is classified as similar to an L1 counterpart while the other L2 sound remains unclassified. It was found that Japanese listeners tend to perceive the English /m/ sound as their native sound /m/, but they do not classify the English /ŋ/ sound when it is in final syllable positions (Aoyama, 2003). Finally, in cases where two L2 sounds cannot be associated with any L1 sound since both are categorised as non-speech categories, the ability to differentiate between them is expected to range from very good to good. In this case, "**non-assimilable (NA)**" refers to non-speech categories. Figure 2.1 below provides a summary of PAM pairwise assimilation types and their associated prediction discrimination.

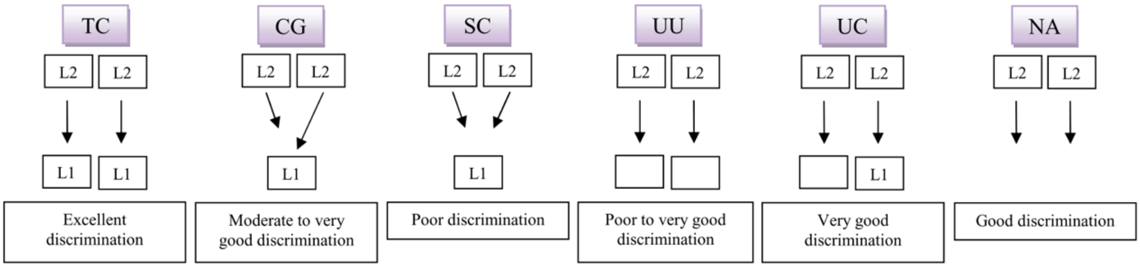


Figure 2.1 Summary of PAM's paired assimilation patterns and their predicted discrimination, according to the similarities between L1 and L2 phonemes (Almbark 2012, p. 47).

Adults L2 learners

The application of PAM has been broadened to include L2 learners (instead of monolinguals and learners with limited L2 input). This expansion was introduced through establishing the PAM-L2 Model by Best and Tyler (2007). The authors speculate that there are notable distinctions among naïve listeners, bilinguals, and L2 learners due to the variations in their exposure to and experience with the second language. PAM-L2 focuses on natural speech interaction in contrast to the controlled laboratory conditions of the original PAM. The model proposes that learners acquire L2 speech sounds at a physical (phonetic) level and then at a more mentally organised level (phonological). Moreover, it claims that the formation of new mental sound categories is influenced by the degree to which the sound patterns in an L2 differ from those in the native language. There are four cases regarding how learners perceive contrasts in L2 speech sound within the PAM-L2 framework:

- **UC or TC contrast**: An L2 category is assimilated into an L1 phonological category. The expectation is that learners will perceive them to be phonologically and phonetically similar. They are also likely to distinguish the L2 contrast member.

- **The CG contrast**: both components of an L2 contrast are associated with a singular L1 sound. However, one of the sounds diverges more—being less typical—while the other is closer to the native category. In this case, a new phonetic and phonological category will be created for L2 sounds that do not neatly align with the L1 category.

- **SC contrast** occurs when two sounds from an L2 contrast are identified as belonging to the same native category. According to PAM-L2, the degree of ease or difficulty in perceiving differences is reliant upon whether the sounds are perceived as excellent or poor representations of the L1 category.

- **UU contrast** happens when two members of an L2 contrast do not correspond to any particular sound category in the learner's native language. In order to accommodate these unfamiliar L2 sounds, it is anticipated that learners will create new phonological categories within their cognitive framework.

Many investigations examined and strongly supported PAM assumption (e.g., Ingram & Park, 1997; Lengeris & Hazan, 2007; Almbark, 2012). For example, Almbark (2012) used acoustic analysis to determine how Syrian Arabic L2 learners produced English vowel contrasts compared to SSBE speakers based on PAM-L2 predictions. The study involved 15 Syrian Arabic L2 participants, 5 females and 10 males, and two SSBE speakers, all of whom pronounced target vowels in the /hVd/ context. The English contrasts (/æ/ -/ɑː/) and (/ɒ/- /ʌ/) were categorised as TC, (/eɪ/- /ɛə/) as SC, and (/ɪ/-/e/) as CG. The findings showed that the PAM TC vowels (/æ/- /ɑː/) and (/ɒ/-/ʌ/) were associated with two discrete native sounds, namely (/aː/-[aːˤ]) and ([o]- /a/), respectively. Even though English contrasts (/eɪ/-/ɛə/) were associated with the single Syrian Arabic sound /eː/ (SC), their production differed phonetically from the native /eː/ sound. Regarding the CG contrast (/ɪ/- /e/), learners produced /ɪ/ closer to SSBE production than /e/. It was thus concluded that the patterns of PAM perceptual assimilation were applicable and could be used to infer how Syrian Arabic L2 speakers produce English vowel contrasts.

Ingram and Park (1997) examined PAM's predictions by examining how Japanese and Korean adult learners distinguish the Australian English vowel contrast (/e/-/æ/) in light of their distinct L1 phonological systems. Japanese emphasises vowel duration for sound distinction, whereas Korean does not. Thus, it was speculated that Japanese learners may utilise their knowledge of vowel duration to comprehend this English distinction, while Koreans may not. Regarding vowel quality, Japanese lacks a sound comparable to Australian vowel /æ/, whereas Korean has an (/e/-/ɛ/) distinction, with /ɛ/ believed to correspond to Australian /æ/. Using PAM's framework, the English (/e/-/æ/) was hypothesised to represent a TC pair for Koreans and a UC pair for Japanese. For both groups, the English contrast was anticipated to be distinguishable. Participants listened to Australian English vowels and selected one of five options to identify the vowel. They then transcribed it using Japanese Kana or Korean Hangul and judged its similarity to their native vowel. The English (/e/-/æ/) contrast presented difficulty for both Japanese and Korean participants. Japanese participants performed better in distinguishing vowel contrasts than their Korean counterparts, even without a direct Japanese equivalent to the Australian /æ/ sound. This finding contrasts with the predictions of PAM. However, the researchers attributed this result to the fact that Japanese learners utilised durational differences between Australian (/e/-/æ/) vowels (due to the L1 transfer effect) that Koreans lacked.

2.4.3 Speech Learning Model (SLM)

Flege (1995) introduced the SLM, which focuses on how native language influences learners' perception of L2 speech patterns. The model concentrates primarily on bilingual immigrants and those who acquired their L2 in a natural environment (with substitutional L2 input) (Flege & MacKay, 2010). The model emphasises the development of language-specific phonetic categories and the motor rules for their production. It operates within a three-level perception-production framework, where information flows from sensory motor processing to phonetic categories, and finally to lexicon-phonological representations. Phonetic categories serve two key functions: defining articulatory goals for speech production and enabling access to speech segments for word recognition. While listeners are typically unaware of these categories during speech processing, they are detailed enough to detect nonnative speech quickly. Additionally, phonetic-level differences can be perceived when attention is focused on them. Following is a summary of the fundamental postulates and hypotheses of the SLM (Flege,1995).

1. Age factor and impact of L2 Experience

Learners maintain their ability to acquire L2 speech sounds throughout their lifetime. It means that the mechanisms involved in L1 development are continuously accessible and can be used for L2 speech learning. Initially, it may be difficult for learners to distinguish between L1 and L2 sounds. However, the increased exposure to the L2 can lead to more refined and accurate sound categories. L2 experience, referring to the cumulative speech input learners gain through verbal communication, particularly in face-to-face interactions, substantially impacts learners' perception and production of L2 sounds. This viewpoint challenges the basis of Lenneberg's (1967) CPH, according to which there is a critical period after which it is impossible to attain native-like sound proficiency due to the loss of neural plasticity. Nonetheless, the SLM emphasises the effect of age on the acquisition of L2 speech sounds and observes that as learners age, their L1 experiences have a greater impact on L2 sounds. Consequently, their ability to perceive and produce L2 sounds may decline and become more challenging as they grow older. Notably, this is not attributable to cognitive decline but to the increased establishment of L1 phonetic categories (i.e., L1 categories become too intense when age increases).

2. Position-sensitive allophones

The SLM suggests that the mapping of L2 sounds to L1 sounds occurs at the level of allophones, which are variations of a phoneme that occur depending on their position in a word, rather than at the level of phonemes themselves. This is based on the understanding that the distribution and realisation of allophones can vary significantly both across languages and within a single language, depending on their position within a word. Consequently, learning to perceive or produce an L2 sound in one position does not guarantee the ability to do so in other positions. For instance, Iverson et al. (2005) discovered that training native Japanese speakers to identify /r/ and /l/ in the initial position improved their accuracy for liquids in that specific context, but it did not enhance their identification of liquids in medial positions or in initial clusters.

3. Interaction between L1 and L2 sound systems

The proposed model postulates L2 learners differ from monolingual native speakers due to interactions between the L1 and L2 phonetic subsystems. Learners' perception and production of sounds are influenced by the phonetic similarities and differences between their native and target languages. Both languages have a mutual influence on each other in this regard. The degree to which each language might affect the other is related to language dominance. According to Flege (2002), late bilinguals may experience a more considerable influence of their L1 on their L2, whereas early bilinguals may experience a more significant influence of L2 on their L1. The model suggests that L1 and L2 sound categories coexist in a shared perceptual space “common phonetic space” and are allophonically related. Learners, however, attempt to divide this space to distinguish between the L1 and L2 categories. Additionally, it suggests that the ease of learning a sound in L2 is determined by the perceived similarity between that sound and a sound in L1. This means that the sounds in the L2 do not exist in isolation but interact with the sounds of the L1. While it is possible for learners to establish new L2 categories, achieving native-like production is not guaranteed because L1 and L2 categories reside in the same phonological domain. The coexistence of phonological systems often leads learners to blend their L1 and L2 categories, leading to the phenomenon of category assimilation. Learners who possess this composite L1-L2 categorisation not only have difficulty achieving native-level proficiency but also differ from monolingual speakers of either the L1 or L2 language. This may provide a framework for understanding why native-like performance is elusive for many L2 learners; therefore, it may not be a feasible goal.

4. Categories, Not Contrasts

The SLM emphasises the significance of individual sounds within the L1 and L2 phonetic subsystems of L2 learners, rather than focusing on contrasts between sound pairs. This focus on position-sensitive allophones is based on the idea that listeners match the characteristics of an incoming sound to a stored representation in long-term memory. This process is crucial because, during real-time speech processing, it is inefficient to eliminate multiple alternative sound candidates. Since categorising speech sounds is fundamental to speech perception in monolinguals, this principle also applies to L2 sound perception, given the perceptual links between L1 and L2 sounds (Flege & Bohn, 2021). Therefore, the SLM provides a robust framework for understanding how L2 learners perceive individual vowels, which is the focus of this thesis.

5. Classification of L2 sounds:

The SLM hypothesises that learners subconsciously and automatically associate L2 sounds with L1 phonetic categories, and that the greater the perceived phonetic dissimilarity between an L2 sound and its closest L1 counterpart, the more likely it is that a new phonetic category will be formed for the L2 sound. The model categories the relationships between L1 and L2 sounds based on their degree of similarity as:

New: In relation to L2 sounds significantly different from any L1 sound, learners may face challenges during the initial stages as they might mistakenly perceive the L2 sound as an L1 sound. However, once they have been exposed to sufficient L2 sounds (increased experience in the L2), they may develop a novel (new) phonetic category. It is predicted that learners will ultimately encounter no difficulties if the sounds in L1 and L2 differ. According to Flege (1995), learners can create a new L2 sound category when they can decode (hear) common phonetic features on multiple sound examples while ignoring the unique ways each speaker may pronounce it. A proficient Arabic learner of English, for instance, can perceive English vowels regardless of speakers, accents, tone, and pitch variations. In addition, Flege discussed that L2 learners must differentiate multiple occurrences of a category from multiple occurrences of other categories while disregarding irrelevant similarities, for example, distinguishing between /t/ and /d/ based on their voicing while neglecting their common place of articulation.

Identical: The SLM posits a mechanism of perceptual equivalence, wherein an L2 sound is perceived as being identical to an L1 sound. Due to the equivalence classification, it is improbable that a new category will be developed, resulting in imprecise production of the L2 sound. This classification aligns with the single-category assimilation pattern of PAM. In addition, according to SLM, when L1 and L2 sounds are perceived as identical, the phonological systems of both languages are affected, and the production of both sounds becomes similar (diaphones).

Similar: This applies to L2 sounds that are not identical to any L1 sound but are close enough (similar) to be categorised or grouped under a pre-existing L1 category. Difficulties appear with similar L2 sounds since they are paired with or assimilated to the nearest L1 category despite the absence of a perfect match. Learners would consider L2 sounds equivalent to their L1 sounds because they have already established phonetic patterns in their native language. Creating a new sound category requires perceiving the distinctions between the L1 and L2 sounds. It is less complicated for adult learners to create a new sound category when they perceive the L2 sound to be similar but still distinguishable from an L1 sound.

In short, the categorisation of sound relationships between L1 and L2 in Flege's SLM is based on the degree of their similarity. It is anticipated that the effects of equivalence classification of similar or identical sounds will impede or prevent the formation of new categories. Nevertheless, as the perceived disparity between the L1 and L2 categories increases, there is a higher probability of forming a novel category. In other words, sounds with a high degree of similarity are categorised by equivalence, whereas those with a lesser degree of similarity lead to creating new categories.

6. L2 Phonetic category formation:

While the SLM explored whether highly experienced L2 learners could achieve native-like proficiency in L2 phonetic categories similar to monolingual native speakers, the model suggests that very few, if any, learners actually reach this level of perfection. Most L2 learners form phonetic categories for L2 sounds that frequently differ from those of native speakers, not because of reduced learning capacity, but due to variations in phonetic input and exposure. These differences arise from factors such as less extensive phonetic input compared to monolingual children, exposure to various L2 dialects, and interactions between L1 and L2 sounds within a shared "phonetic space". Additionally, the unique features of L2 sounds, particularly those do not present in L1, and the differing importance of features between L1 and L2 sounds, further contribute to these variations.

The SLM posits that L2 learners form new phonetic categories similarly to how infants develop categories in their L1, through the creation of auditory equivalence classes based on the statistical properties of sounds they hear. However, L2 learners face the added challenge of overcoming L1 influences, requiring them to discern differences between L1 and L2 sounds. The process of forming these categories involves learning language-specific cue weighting, which can vary significantly across languages. The ability to form new L2 phonetic categories depends on the perceived dissimilarity between L1 and L2 sounds and the age of first L2 exposure, with younger learners more likely to form new categories. However, the diverse and varied input L2 learners receive complicates this process, potentially making it as long or longer than L1 category formation. Reliable measures of cross-language dissimilarity are necessary to accurately test SLM predictions (Flege & Bohn, 2021).

7. Perception before production

The SLM challenges the common belief that the foreign accents and specific errors in vowels and consonants observed in L2 learners are primarily due to age-related declines in the ability to learn new articulatory forms. Instead, the model argues that these errors often stem from perceptual issues, emphasising that accurate perception is crucial for accurate production. The model proposes that L2 learners develop perceptual phonetic categories based on what they hear, which in turn guide their motor skills, or realisation rules, for producing sounds. As learners progressively refine their perception of L2 sounds, their production should increasingly align with these perceptual categories. However, the SLM does not specify a timeline for this alignment, recognising it as a gradual and individualised process.

Support for SLM assumptions has been documented in various studies (e.g., Flege, 1987; Bohn & Flege, 1992, among others). For example, Flege (1987) analysed the production of the French vowels /y/ and /u/ by American learners. Their acoustic features were compared to those of French-native speakers. The English learners were divided into three experience levels: beginners with no L2 exposure, graduate students with a one-year stay in France, and advanced university students. Two groups of learners (with L2 experience) were able to produce the French phoneme /y/, which has no English equivalent, with F2 values that fell within the range of values measured in native French speakers. However, English learners articulated the French vowel /u/ with an accent (exhibiting higher F2 values than native French speakers). This demonstrates that because French /u/ matches its English counterpart (similar phoneme), L2 learners transferred the acoustic properties of their native language (English) to the French vowel.

Likewise, Bohn and Flege (1992) proposed that while L2 experience would not affect the production of familiar English vowels like /i/, /ɪ/, and /e/ by German learners, it could improve their articulation of the unfamiliar vowel /æ/ in a manner that closely resembles that of native speakers. To validate the SLM hypothesis, they carried out an acoustic analysis to compare the vowel production of experienced and inexperienced German learners to that of native L1 English speakers. Each of these groups consisted of ten participants with an equal representation of genders. The results showed that neither group differed in the acoustic properties of their productions of similar sounds. In contrast, the experienced group produced the new vowel /æ/ close to L1 English productions, which supports the SLM. In addition, the study conducted a vowel intelligibility task in which three American listeners were asked to distinguish between the two German groups' productions. They were given /bVt/ sound tokens and asked to determine the word they believed to have heard from a list of six. Contrary to the model's predictions, the results revealed no significant difference between the two German groups' clarity of their /æ/ vowel production. The researchers speculated that the discrepancies between the two experiments may have been caused by using different criteria for analysing the productions or by the fact that both German groups had similar, shorter productions of the /æ/ vowel, resulting in comparable listener interpretations. These results may indicate that just because a non-native production is identified as such through acoustic evaluation does not inherently make it less intelligible to L1 listeners.

2.4.4 The Revised Speech Learning Model (SLM- r)

Similar to its predecessor, the Revised Speech Learning Model (SLM-r) (Flege & Bohn, 2021) explores how sequential bilinguals produce and perceive position-sensitive allophones of L2 vowels and consonants. It seeks to explain the lifelong evolution of phonetic systems in response to phonetic input encountered during natural L2 acquisition. While the SLM-r retains several core elements from the original SLM (Flege, 1995), it also introduces new concepts, informed by numerous published studies. The fundamental principles of the SLM-r are summarised as follows:

1. **L2 input and experience**

   The original SLM focused on whether highly experienced L2 learners could achieve native-like proficiency in L2 phonetic categories, comparable to monolingual native speakers. However, the SLM-r has moved away from this idea, recognising that perfect mastery is unlikely due to the interaction between a bilingual's L1 and L2 phonetic systems and the differing phonetic input L2 learners receive compared to native speakers. The SLM-r acknowledges that these factors prevent L2 learners from perfectly matching the phonetic proficiency of monolingual native speakers.

Although the SLM suggests that L2 learners gradually discern L1–L2 phonetic differences with experience, leading to new L2-specific categories, it lacks a clear method for measuring how phonetic information accumulates or how much input is needed for this process. Flege and Bohn (2021) argue that relying on Length of Residence (LOR) as a metric is problematic, as it does not accurately reflect the quantity or quality of phonetic input learners receive. Consequently, the SLM-r emphasises the importance of both the quantity and quality of phonetic input in forming new L2 categories. It defines phonetic input as the sensory experiences associated with L2 sounds in meaningful conversations. To better estimate L2 input, SLM-r introduces the concept of full-time equivalent (FTE) years, calculated by multiplying LOR by the proportion of time the L2 is used, thus providing a more accurate measure than LOR alone. The model also stresses that L2 input quality has been overlooked in research, underscoring the need for more refined measures. Learners who encounter accented L2 input might form distinct phonetic categories from those who receive native-like input. Additionally, the context in which this input is absorbed significantly impacts its retention and application in learning. For example, Arabic learners in FL classrooms taught by local instructors may develop different phonetic categories compared to those studying abroad and interacting with both native and non-native speakers.

2. The category precision hypothesis

The SLM-r replaces the "age" hypothesis with the "L1 category precision" hypothesis, which posits that more precisely defined L1 categories at the time of first L2 exposure enhance the ability to discern phonetic differences between L1 and L2 sounds, leading to the formation of new L2 phonetic categories. According to this hypothesis, individuals with well-defined L1 categories are more adept at recognising differences between L1 and L2 sounds, increasing the possibility of developing new L2 categories. While L1 category precision generally improves through childhood into early adolescence, significant individual differences persist at all ages, separating it from age-related neurocognitive plasticity.

The SLM-r acknowledges that while cross-language phonetic research often highlights differences between languages, significant variations also exist in how individual speakers of the same language produce and perceive phonetic categories. The model presents category precision as an intrinsic quality shaped by individual auditory capabilities rather than external language exposure. It links differences in phonetic category precision to endogenous factors such as auditory acuity, early-stage auditory processing, and auditory working memory, all of which impact sound production and perception.

3. Bilingual phonetic categories

The SLM-r asserts that the capacity to form new phonetic categories remains intact across the lifespan, yet not all L2 sounds will necessarily result in the creation of new categories. Some L2 sounds are so similar to an L1 sound that an L1-for-L2 substitution would go unnoticed by monolingual speakers of the target L2. The formation of new phonetic categories is influenced by factors such as the perceived dissimilarity between L1 and L2 sounds, the precision of L1 categories, and the quality and quantity of L2 input. The process of phonetic category formation is gradual, relying on distributional learning, where learners first discern differences between L1 and L2 sounds, then group similar L2 sounds together, and ultimately delink these from their L1 counterparts to establish distinct categories. The model also posits that the expansion of L2 lexicon may facilitate this process, reinforcing pre-existing weak categories, while L2 sounds that closely resemble L1 sounds may lead to the formation of composite categories rather than distinct new ones.

4. Features weighting

The original SLM proposed that L2 phonetic categories might differ from those of native speakers due to differences in how phonetic features are weighted between L1 and L2. However, Flege and Bohn (2021) argued that this "feature hypothesis" conflicted with the model's core principle that the mechanisms used for learning L1 remain effective throughout life and also apply to L2 learning. As a result, the revised SLM-r adopts the "full access hypothesis," which asserts that L2 learners, even those learning later in life, can access all necessary phonetic features to form L2 categories, using the same processes as in L1 learning. Additionally, SLM-r proposes that both new L2 categories and composite L1-L2 categories are shaped by input distributions, with the weighting of perceptual cues being influenced by their reliability in facilitating accurate and rapid categorisation.

5. Individual differences in speech learning ability

The SLM-r highlights the critical role of individual differences in L2 speech learning. It attributes intersubject variability in L2 sound production and perception to factors including how L1 phonetic categories were defined when learners first encountered L2 sounds, their perceptual linkage of L2 to L1 sounds through interlingual identification, and the perceived dissimilarity between L2 and L1 sounds. Additionally, inherent speech learning abilities, auditory processing skills, phonological short-term memory, and the quantity and quality of L2 phonetic input received play crucial roles. Enhanced auditory acuity, early-stage processing, working memory, and abilities like mimicry, musical training, selective attention, and phonemic coding further shape L2 learning outcomes, highlighting the need to consider these individual differences to fully understand the complexities of L2 acquisition.

6. Perception and production coevolve

The SLM posits that perceptual representations set an upper limit on the accuracy of L2 sound production, suggesting that accurate perception is a prerequisite for accurate production. In contrast, the SLM-r proposes that production and perception coevolve in a bidirectional relationship, where improvements in one can enhance the other, and both develop simultaneously. While the SLM emphasises a sequential process with perception preceding production, the SLM-r advocates for a more dynamic, interactive development of these skills, highlighting their mutual influence throughout the learning process.

7. **L2 Speech Learning Milestones**

The SLM-r shifts the focus from group-based comparisons to individual learner variability in L2 speech learning. Moving away from the original SLM, which categorised learners broadly as "early" or "late," the SLM-r points out that averages across groups can hide essential individual variations. It t promotes viewing each learner as a distinct case study, emphasising the need for detailed, personalised data to accurately understand their unique learning trajectories and milestones. This approach highlights the importance of detailed, individualised data and calls for new research methods to explore the factors influencing each learner's progress in acquiring L2 speech sounds.

2.4.5. Second Language Linguistic Perception Model – L2LP (Escudero, 2005)

Introduced by Escudero (2005), the L2LP model is a derivative of the original SLM and PAM models. It comprehensively explains the production and perception of cross-linguistic speech and includes all phases of L2 learning. While the L2LP model recognises the effect of age, it suggests that high-quality L2 input can counterbalance the reduced adaptability observed in the adult brain. L2LP acknowledges that carrying out separate analyses of L1 and L2 categories can provide valuable insights into the challenges and development experienced by L2 learners. The L2LP model addresses four primary aspects of the acquisition of L2 sound:

- **Initial State:** Learners initially use a copy of their native language sound system to comprehend L2 sounds. They use their L1 perceptual parameters (familiar properties) to decode and interpret the acoustic qualities of L2 sounds, transforming them into recognisable speech patterns. That is to say, L1 acquisition plays a crucial part in the initial stages of L2 learning, as learners initially produce and perceive L2 sounds in a manner analogous to their L1 system.

- **Learning task:** The purpose of the learning task is to acquire an optimal (best) perception of the L2 sound system. There are two different kinds of tasks that can be identified: those that focus on perception and those that emphasise representation. The perceptual tasks involve the process of auditory perception and sound processing. Learners may need to adjust or even develop new ways to "hear" and interpret distinctions in sounds that were not present or deemed significant in their native language. For example, Spanish speakers learning the English contrast /i/- /ɪ/ must adjust their perceptual processing (using the F1 parameter) in order to discriminate between the two sounds. Moreover, they may need to learn to focus on the new auditory cue (duration), which they may not have previously attended in their L1. Thus, perceptual tasks necessitate modifying or creating novel auditory perception methodologies. On the other hand, the representational task pertains to mentally categorising or representing sounds. Once learners have developed the ability to discriminate between sounds perceptually, they must subsequently learn how to classify them cognitively according to the L2 system. For instance, despite the ability of Spanish speakers to discriminate between the phonemes /i/ and /ɪ/, they might still mentally classify them as the same sound due to the presence of a single category for that sound /i/ in their L1 system. To prevent ambiguity in the lexical representation of English words containing the vowels /i/ and /ɪ/, Spanish learners need to develop a new category for /ɪ/.

- **Development**: In this stage, learners might form new L2 phonological representations. They employ fundamental learning principles from Universal Grammar (UG), including distributional learning and the Gradual Learning Algorithm (GLA), which are also used when acquiring the L1. Distributional learning facilitates the formation of categories for L2 features, such as duration cues in Spanish learners. In contrast, the GLA allows for the fine-tuning of L2 categories.

- **End state**: At this phase, learners might or might not attain a perfect (optimal) perception of L2 sound categories. Their ability to distinguish L2 sounds depends on their exposure to the language and the quality of the L2 input they hear. Consequently, the model does not specify definitive outcomes for L2 learners but instead emphasises the significance of high-quality L2 input over potential cognitive adjustments in adulthood. Ideally, individuals should possess equal competence in both their L1 and L2 systems, considering that the primary language remains unaffected by the phonetic patterns of the L2.

The L2LP model, like the PAM and SLM, anticipates the degree to which L2 production and perception align with L1 patterns. It identifies the following three scenarios for L2 learners:

- **New scenario**: when two L2 speech sounds are assimilated to a single L1 category (similar to PAM's SC), it is predicted that learners will struggle to differentiate between them, resulting in poor performance. This scenario is common among L2 learners whose L1 inventory contains fewer vowels than the L2 (Escudero 2009), for example, Arabic in comparison to English.

- **Similar scenario:** when two L2 speech sounds are linked with two separate L1 categories (similar to PAM's TC), the discrimination is predicted to be straightforward**.**

- **Subset scenario**: two L2 speech sounds are mapped into two or (multiple) L1 sounds. This scenario may cause difficulties for L2 learners whose L1 has a richer vowel inventory than L2 (Escudero & Boersma, 2002).

Several research studies (e.g., Escudero & Chládková, 2010; Elvin et al., 2016) have documented the L2LP model's validity. In their study, Escudero and Chládková (2010) investigated how Peruvian Spanish listeners perceive vowels in American English (AmE) and SSBE accents. They used principles derived from the PAM and integrated them into the L2LP framework. 40 Spanish listeners heard a total of 205 auditory vowel stimuli, made up of 18 English vowels (from both AmE and SSBE) played 10 times and 5 Spanish vowels played 5 times. They were then instructed to select the corresponding Spanish vowel depicted orthographically on the screen for each sound they heard.

The primary result of this study indicates that Spanish speakers heard American and British English vowels distinctly. Moreover, multiple patterns of discrimination were determined during the analysis: The AmE vowels /ɪ/, /æ/, and /e/ were perceived as the Spanish vowel sound /e/. The AmE contrast /ʊ/-/ɔ/ was perceived as the Spanish vowel /o/. The AmE vowel /ɑ/ was most frequently perceived as Spanish vowel /a/, while /ʌ/ was identified as Spanish vowels /a/ and /o/ with varying frequency. In the context of SSBE vowels, /ɪ/ and /e/ were perceived as Spanish /e/, whereas /æ/ was heard as the Spanish /a/. The SSBE /ɑ/ underwent partial assimilation to the Spanish /o/, while the SSBE contrast /æ/-/ʌ/ was perceived as the Spanish /a/. The vowel contrast /u/-/ʊ/ was observed to correspond to the Spanish /u/. These observations align with the "new scenario" of the L2LP model, indicating that limited discrimination is expected. Most English vowel contrasts were mapped into a single Spanish vowel (except AmE /ʌ/, identified as Spanish vowels /a/ and /o/). Such assimilation challenges are to be expected since Spanish learners possess fewer vowels than AmE or SSBE, in line with Escudero (2009). The following section provides a comprehensive overview of vowel sounds.

2.5 Overview of vowels

Vowels are voiced sounds that, in most languages, constitute the core of syllables and words and are distinguished by their open vocal tract during articulation (Ashby & Maidment, 2005). In contrast to consonants, whose articulation can be identified, vowels require a broad articulatory channel. They are therefore better defined by examining their acoustic characteristics rather than precise articulatory movements (Ladefoged & Johnson, 2015). This could potentially result in discernible differences in production among individuals, as accurately identifying the exact positioning of the tongue can be challenging. Three primary characteristics determine the articulation of vowels: tongue height, tongue position, and lip position. Vowel height is determined by the relative rise of the tongue within the oral cavity, producing high (e.g., /i/), low (e.g., /ɑ/), and mid-position (e.g., /e/) vowels. The location of a vowel is determined by tongue positioning—frontness or backness—during articulation. Front vowels, like /i/, are formed by aligning the front of the tongue with the soft palate, while back vowels, like /u/, involve positioning the tongue's back towards the soft palate. Lip rounding can also distinctly differentiate vowels, as observed in English's rounded /u/ and unrounded /i/ (Roach, 2010; Ashby & Maidment, 2005; Colantoni et al., 2015).

In addition, vowels can be characterised either phonologically or phonetically. The phonological or contrastive nature of vowels refers to how their substitution alters the meaning of a word. In English, the vowels /i/ and /ɪ/ in the words ’sleep’ and ’slip’ respectively are contrastive because they produce distinct meanings within the same phonological context. This phonological aspect elegantly connects to the concept of underlying representation (UR). UR represents the abstract form of a word in a speaker's mental lexicon before any phonological rules are applied. UR, represented by slashes / / in transcription, contains a word's generalised form before it undergoes phonological adjustments. However, vowels in the phonetic sense are described with regard to articulators (e.g., the lips and the jaw) and their acoustic properties (e.g., formant frequencies and duration). In parallel, the surface representation (SR) represents the phonetic realisation of a word, which appears in speech after phonological processes have shaped the UR, represented by square brackets [ ] in transcription.

Vowels are commonly described according to their formant structures, which provide observations into the resonances of the vocal tract and the corresponding articulatory shapes. In this regard, the first and second formant frequencies (F1, F2) are crucial, with F1 values related to vowel openness (height) and F2 values related to vowel frontness (Ladefoged & Disner, 2012; Kent & Read, 2002). Measuring the formant frequency values of F1 and F2 at a stable point in a vowel's sound spectrum (typically the midpoint) has been a commonly employed method for describing monophthong vowels in L1 and L2 acoustics research. (e.g., Munro, 1993; Peterson & Barney, 1952). This traditionally static approach is widely adopted because it is thought to represent the target vowel position a speaker aims for during speech (Peterson & Barney, 1952). However, it is well-documented in a substantial body of research that covers both L1 and L2 context (Harrington & Cassidy, 1994; Hillenbrand & Nearey, 1999; Morrison & Assmann, 2013;Hillenbrand, 2013; Almurashi et al., 2019) that the reduction of vowel acoustic characteristics to a static parameterisation is constrained and inadequately exhaustive for the accurate description of vowels. Alternatively, using the dynamic cues, specifically Vowel

Inherent Spectral Change (VISC) models, not only for diphthongs but also for monophthongs, has proved to be valuable in identifying vowel sounds in a variety of languages and provide crucial information (e.g., spectral movements) that is not captured by simply selecting a single point (mid-point).

In more straightforward terms, researchers have found improved monophthong identification in models incorporating VISC. This integration provides a more thorough depiction, a deeper understanding, and a more accurate representation rather than relying on a single-point measurement. For example, Almurashi et al. (2020) and Almurashi (2022) conducted thorough research on articulating SSBE vowels, specifically on VISC models. They analysed the production patterns of SSBE vowels in Saudi Arabic learners of English, comparing them to their native dialect (Hijazi Arabic) and their second language (SSBE). The acoustic analysis employed not only static measurements (vowel midpoint) but also dynamic measurements (the amount of vowel change 'offset', spectral rate of vowel change 'slope', and the direction of vowel shift 'direction') to capture the inherent dynamic nature of vowels effectively[27]. The research used a word list to examine particular vowels[28] in different consonant contexts. The data collection involved three groups, each consisting of twenty participants, representing Arabic learners of English, Arabic speakers, and English speakers of both genders. The findings (support the dynamic theories of vowels) revealed that dynamic cues provide insights into the production patterns of all three groups that are not typically derived from static measurements alone. It was found that dynamic cues, particularly those represented in the three-point model, demonstrated the highest classification accuracy across all three participant groups.

[27] Additional cues, including vowel duration, fundamental frequency (F0), and the frequency of the third formant (F3), were also subject to investigation.

[28] Arabic speakers produced the Saudi vowels (/i iː a aː u uː oː eː/) in the words pronounced in the phrase "ktoːb _____ marteːn," which means 'Write ___ twice' whereas L2 learners and English speakers articulated the SSBE vowels (iːɪ e ɔː ɑː ʊ uː æ ɒ ʌ) within the word produced in the phrase ‘say___ again'.

Alongside the formant structure, the duration of vowels also plays a significant role in classifying vowels. Vowels can be described according to whether they are long or short, resulting in a vast array of sounds in spoken language. As shown in the figure below, there are distinctions in the duration of the vowels /æ/ and /e/ as produced in the words ‘bet’, ‘bed’, and ‘bad’ (Kent & Read, 2002, p.127).

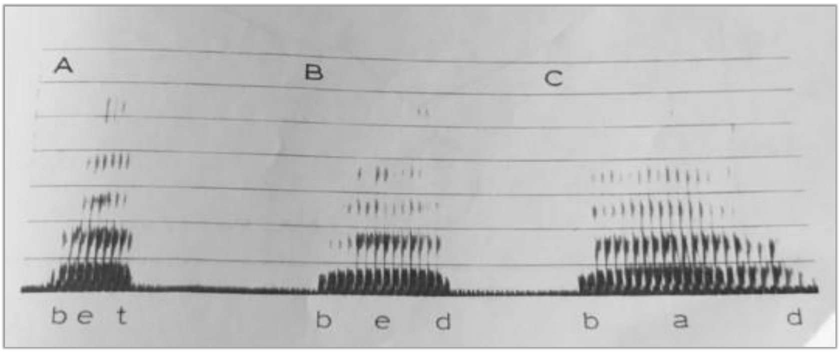


Figure 2.2 *Spectrographic illustration of differences in the length of the* vowels /e/ and /æ/ (Kent & Read, 2002; 127).

Vowels produced with an open jaw (low vowels) have longer durations than vowels produced with a closed jaw position (high vowels) (Hillenbrand et al., 1995). Additionally, vowels differ based on their tense and lax qualities, with tense vowels having longer durations than lax vowels. The tense and lax distinction emphasises the importance of articulatory gestures and muscular involvement in production. Tense vowels require deliberate and well-defined articulatory movements, necessitating increased muscular effort, resulting in lengthier and more distinct vowel sounds. On the other hand, lax vowels are known for their relaxed muscular tension and faster articulatory patterns, which can lead to shorter vowel productions (Chomsky & Halle, 1968; Lehiste & Peterson, 1961). This section presented a brief overview of vowels. The following section provides a review of the participants' L1 (Arabic) and L2 (English) languages, emphasising differences in the quantity and quality of L1 and L2 vowel systems.

2.6 Arabic and English vowel systems

A comprehensive summary discusses Arabic vowels' qualitative and quantitative features (duration, formants), focusing on the Saudi Arabian dialect. Following this is a concise overview of the duration and quality of English vowels, focusing on SSBE, AmE, and Australian English (AusE). Reviewing these vowel systems is essential, as participants received a brief overview during the study's preparation sessions. Comparing the simpler Saudi Arabic vowel system to the more complex English vowel inventories helps participants understand why perceiving English vowels can be challenging and highlights the value of the training. Additionally, understanding the vowel differences among various English accents enables participants to perceive English vowels more accurately in diverse settings, a crucial skill for effective real-world communication. This knowledge also potentially sharpens their ability to discern subtle differences between vowels across different English accents.

As vowels play a significant role in accent variation, different symbols frequently represent a vowel sound in the same lexical item across various dialects. For instance, the vowel in the word ‘hot’ is /ɔ/ in most varieties of AusE, /ɑ/ in AmE, and as /ɒ/ in SSBE. This diversity poses difficulties when comparing dialects (Cox & Fletcher, 2017). Therefore, linguists frequently employ lexical sets as a standard reference to address this issue. In his work on Accents of English (1982), John Wells devised lexical sets in order to avoid confusion when discussing vowel sounds. This has enabled studies to examine variations in the phonological inventory and phonetic variation of many English accents.

2.6.1 Arabic vowels

Arabic, categorised as Semitic (Hetzron, 1997), includes standard and dialectal varieties. Modern Standard Arabic (MSA) is derived from Classical Arabic (CA) and is primarily used in official contexts such as education, media, and political dialogues (Holes, 2004) due to its more extensive vocabulary and more straightforward grammar than CA, which is primarily used in religious contexts (Huthaily, 2003). MSA or CA employs a three-quality vowel system with short/long distinctions (/a aː, u uː, i iː/). Thus, the short vowels (/a, u, i/) and their long equivalents (/aː, uː, iː/) define the monophthongs vowels of MSA and CA (Newman & Verhoeven, 2002; Mitchell, 1993; Al-Ani, 1978). The representation of short vowels in the script is achieved through the use of diacritic markers. The following diacritics represent Arabic short vowels and are positioned either above or below the root consonants:

Fathah ◌َ (i.e., /a/): it is written above the letter
Dammah ◌ُ (i.e., /u/): it is written above the letter
Kasrah ◌ِ (i.e., /i/): it is written underneath the letter

The following table presents lexicons of MSA short vowels:

MSA/CA vowel	IPA	MSA/CA word	English gloss
/a/	/batˤal/	بَطَل	hero
	/darasa/	دَرَسَ	studied
/u/	/burʒ/	بُرج	towel
	/ʤud/	جُد	diligence
/i/	/bint/	بِنت	girl
	/bin/	بِن	son

Table 2.1 Examples of Arabic short vowels.

The long vowels are depicted using the letters provided below, and lexicon examples can be found in Table 2.2 below:

1- أ /aː/
2- و /uː/
3- ي / iː /

MSA/CA vowel	IPA	MSA/CA word	English gloss
aː	/saːr/	سار	to walk
	/kitaːb/	كِتَاب	book
uː	/suːq/	سوق	market
	/kasuːl/	كَسُول	lazy
iː	/fiːl/	فيل	elephant
	/kariːm/	كَرِيم	generous

Table 2.2 Examples of Arabic long vowels.

It is important to note that Arabic dialects do not uniformly conform to the vowel sounds of MSA/CA. Some dialects include additional vowel sounds like central (/ə/) or mid (/e/, /o/) vowels. For example, the Syrian dialect features 11 vowels— (/a aː u uː i iː e eː o oː ə/) (Almbark & Hellmuth, 2015), Jordanian dialect uses eight—(/a aː u uː i iː eː oː), while the Moroccan dialect includes five—(/aː uː iː ʊ ə/) (Al-Tamimi, 2007). This shows that the range of vowel sounds in different Arabic dialects varies not only from the MSA but also from one dialect to another (Newman, 2002).

Regarding diphthongs, Standard Arabic (MSA, CA) includes two: /aj/ and /aw/. These are maintained in particular dialects of Yemen and some peninsula dialects. However, in various Arabic dialect variants, like those spoken in Saudi Arabia, Syria, and Cairo, they have become monophthongs /eː/ and /oː/ instead (Almbark & Hellmuth, 2015; Watson, 2002; Abdoh, 2011). As an illustration, the word "house" is realised as /bajt/ in MSA, whereas "voice" is realised as /ṣˤawt/. However, in the Saudi dialect, these vowels are monophthongal: "house" is realised as /beːt/ and "voice" as /sˤoːt/.

Syllables in Arabic cannot start with a vowel but are introduced by a glottal stop in the absence of a supraglottal consonant; this is referred to as Hamza "ء" (Kopczynski & Meliani, 1993). Apart from the dialectal variation discussed above, Arabic vowels demonstrate variable context-dependent variation. In particular, vowels next to emphatic consonants like /ṭ/, /ḍ/, /ð̣/, and /q/ can be subject to retraction, lowering, and/or rounding (Jongman et al., 2007).

2.6.1.1 Vowel quantity and quality

In regard to the quality of Arabic vowels, there is a noticeable distinction between short and long forms. Several studies (e.g., Alghamdi, 1998; Abou Haidar, 1994, among others) have examined Arabic vowels' production in different Arabic dialects. For instance, Alghamdi (1998) investigated whether the vowels of MSA (/a aː u uː i iː/) have formant values comparable to those of other Arabic dialects (Saudi, Egyptian, Sudanese). The study engaged 15 male Arabic speakers (5 per dialect) and presented words in CVC syllables, each containing one of six Arabic vowels. To avoid potential context effects, stimuli were presented in isolation. While there were no significant differences between short and long vowel durations, there were differences in the first and second formant frequencies (the centralisation of short vowels was greater than that of their longer counterparts). In addition, Arabic vowels produced by individuals of the three dialects differed significantly from MSA. For instance, F1 in /u/, /uː/, and /i/ were substantially higher for Saudi speakers than for Egyptian and Sudanese. In the Egyptian dialect, the vowel /aː/ was considerably lower than its equivalent in Sudanese and Saudi. However, the formant values for each dialect could have exhibited more variation if the researcher had used target words in carrier phrases or natural speech. Jurafsky and Martin (2009) suggest that carrier phrases are more effective and natural than isolated words when examining specific segments.

In a related vein, Abou Haidar (1994) identified substantial discrepancies in formant values across various Arabic dialects, including Lebanese, Jordanian, Syrian, Saudi, Qatari, Tunisian, Emirati, and Sudanese, presenting target words in CVC monosyllabic format. The study recruited eight speakers, each representing a different dialect. One of the primary limitations of this study is its small sample size: selecting one individual to represent each dialect may not provide an accurate representation. Nevertheless, regarding the results, both Abou Haidar's (1994) and Alghamdi's (1998) studies signal that the quality of Arabic vowels is subject to variation based on the spoken dialect. This suggests that Arabic learners' difficulties with English vowels may vary according to their dialect (Newman & Verhoeven, 2002; Alotaibi, 2018). As an illustration, (Alotaibi, 2018) found that Arabic learners from two distinct dialects (Saudi and Tunisian) showed significant differences in their ability to discriminate American vowel contrasts (/æ /-/ʌ/, /i/-/ɪ/, /u/-/ʊ/).

The length and quantity of Arabic vowels hold significant importance in Arabic (Newman & Verhoeven, 2002). Through their research on Lebanese Arabic, Khattab (2007) and Khattab & Al-Tamimi (2008) determined that the duration of long vowels is approximately double that of short vowels. Scholars typically regard Arabic as a quantitative language whose speakers rely significantly on vowel duration to create Arabic contrasts (e.g., Mitchell, 1993; Munro, 1993; Khattab, 2007). This might cause difficulties for Arabic learners when producing and perceiving English lax/tense vowels (Cebrian, 2006; Munro, 1993). For instance, Munro (1993) explored the production of American vowels by 23 L2 speakers (21 males and 2 females) from various Arabic dialects, including Saudi Arabic, Kuwaiti, Jordanian, Egyptian, Syrian, Sudanese, and Palestinian. He found that Arabic learners commonly confuse the tense/lax vowels of American English with the Arabic long and short vowels, respectively.

Concerning the tenseness and laxity of Arabic vowels, a prevalent debate exists. Numerous contemporary studies (e.g., Almurashi et al. 2019; 2020, Almbark & Hellmuth 2015; Khattab & Al-Tamimi 2008) have found that there is a distinction between them in terms of both quality and quantity. For example, a study conducted by Almurashi et al. (2019, 2020) found that the duration of vowels in Hijazi Arabic (HA)[29] plays a significant role in vowel classification; however, the effectiveness of formant frequencies in identifying vowels is unavoidable. The quantity alone was determined to be insufficient in distinguishing all HV vowels. Therefore, the significance of Arabic vowels' quality and quantity is nearly equivalent. Similar conclusions were reached by Almbark and Hellmuth (2015), who investigated the acoustic correlates of long/short vowel contrasts in Syrian Arabic. Their research, which involved 15 Syrian speakers who produced a complete set of vowel categories

[29] A dialect of Arabic, predominantly spoken in the western region of Saudi Arabia, specifically in Jeddah, Medina, Taif, and Makkah

in a /hVd/ context, revealed a distinction in Syrian Arabic between long and short vowels both in terms of quantity and quality. This section offered an overview of the literature concerning Arabic vowel systems. As Saudi Arabic learners are the focus of this thesis, the following section provides a general description of the SA vowel system.

2.6.1.2 Saudi Arabic vowel system

Saudi Arabia is located at the centre of the Arabian Peninsula. Saudi Arabia's official language is Arabic, and its diverse regions feature a variety of dialects, including the Northwestern, Hijazi, Najdi, Southern, Eastern, Asir, and Jizan variations. The vowel system of SA comprises the same three short vowels from the MSA/ CA Arabic (/u a i/) as well as their corresponding long counterparts (/uː aː iː/). In addition, it includes the two long mid vowels (/eː oː/), which originated from the MSA/CA Arabic diphthongs /aw/ and /aj/, as explained in the preceding section. This accounts for a total of eight monophthongs within the SA vowel system, as supported by previous research (Algethami, 2023; Almurashi et al., 2029, 2020; Abdoh, 2011; Jarrah, 1993). While the majority of these referenced studies primarily focus on the HA dialect, it is anticipated that participants in this thesis who belong to the northwest region (lived in Alwajh City) will likewise manifest the same eight monophthong vowels.

In a recent study carried out by Algethami (2023), eight vowels in SA dialect, namely /u a i uː aː iː eː oː/, were reported, although the study did not specify the dialects of the SA participants involved (see Table 2.6 for keyword examples). These vowels were produced by four male SA learners studying in the United Kingdom. Almurashi's (2022) study consistently confirmed the existence of eight SA vowels, but he explicitly specified that the recruited SA subjects speak the HA dialect. The SA vowels were collected from a group of twenty intermediate English language learners studying in Saudi Arabia. Even with a modest sample size and unspecified dialect focus in Algethami's (2023) research, the findings align with those found by Almurashi et al. (2023) and with previous studies considering Hijazi dialects (e.g., Abdoh, 2011; Jarrah, 1993; Mousa, 1994) confirming the existence of eight monophthong vowels in the Saudi Arabic vowel system. The table below presents the SA vowels along with relevant keywords.

SA vowel	IPA	SA word	English gloss
/a/	/baħar/	بَحَر	sea
	/ʃatˤtˤ/	شَط	shore
/aː/	/ɣaːs/	غاص	dived, dove
	/naːr/	نار	fire
/u/	/duʃʃ/	دُش	a bath
	/ħutˤtˤ/	حُط	put
/uː/	/ħuːt/	حوت	whale
	/nuːr/	نور	light
/i/	/qirʃ/	قِرش	shark
	/ʕirs/	عِرس	wedding
/iː/	/tˤiːn/	طين	mud
	/tiːn/	تين	fig
/eː/	/xeːtˤ/	خيط	fishing line
	/sˤeːd/	صيد	hunt
/oː/	/moːj/	موج	wave
	/boːt/	بوت	boat

Table 2.3 A list of SA vowel sounds, along with examples for each vowel.

When considering the analysis of the acoustic attributes of the Saudi vowel system, Almurashi (2022) is a noteworthy research study for its comprehensive consideration of both static and dynamic cues. As said previously, the focus of his study was on HA vowels /u a i uː aː iː eː oː/. The results revealed significant differences between long and short vowels when using static or dynamic models. One of the main findings of this study was that identifying monophthong vowels based solely on the mid-point of vowels was insufficient and simplistic. That is, while the static approach (vowel's midpoint) did not yield statistically significant disparities in the production of particular vowel pairs, a closer examination using dynamic approaches revealed distinctions within the same vowel pairs.

The figure below depicts the visual presentation of HA vowel data from Almurashi's (2022) study, taking the static model (the vowel midpoint) into account. A distinct separation was observed between the normalised F1 and F2 values in the z-scores for the majority of HA vowels. The examination of the vowel space in Figure 2.3 revealed that short vowels occupy a more central position in comparison to their long counterparts. Still, there were cases where the vowel sounds overlapped, specifically between /oː/ and the vowel pair /uː/ and /u/, as well as between /i/ and /eː/. When the research employed the dynamic cues, particularly the direction model, a similar pattern was observed: short HA vowels exhibited distinct directional characteristics from their long counterparts. It was concluded that Arabic vowels, specifically short and long HA vowels, show differences not only in terms of quantity but also in terms of quality. This finding aligns with expectations, as it is widely recognised that articulatory duration and tenseness often display interrelated features (Almurashi et al., 2020; 2019).

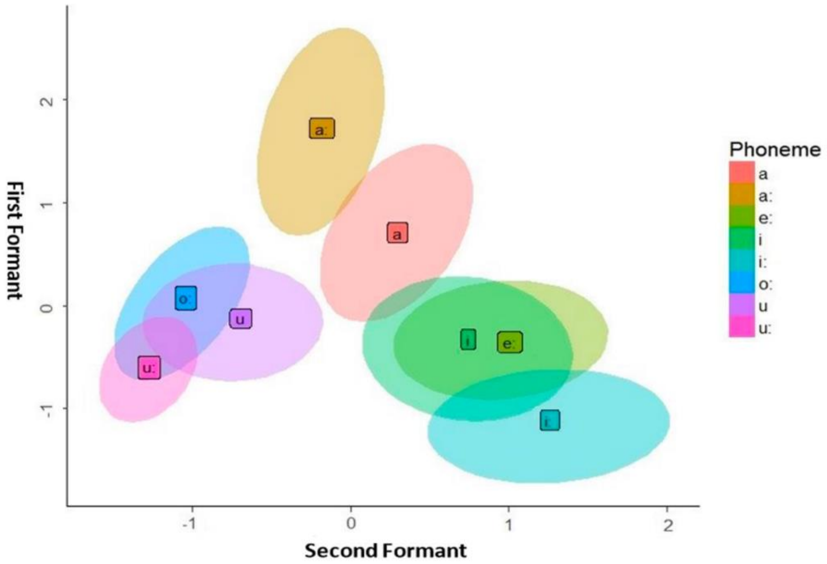


Figure 2.3 A scatter plot illustrating the normalised F1 and F2 midpoint values for Saudi Arabic vowels in the Hijazi dialect, sourced from Almurashi (2022).

In relation to the other cues (duration, F3, F0), Almurashi (2022) found that vowel duration played a substantial role in the classification accuracy for HA vowels, exceeding the influence of the third formant frequency and fundamental frequency. This result is unsurprising, as previous research has demonstrated that Arabic is a language that places a significant emphasis on quantity, with its speakers relying heavily on vowel duration when distinguishing Arabic distinctions (see, for example, Khattab, 2007 and Munro, 1993). The following section provides a general review of the English vowel system, with a specific focus on the standardised variations, including SSBE, AmE, and AusE.

2.6.2 English vowels

English (in all its varieties) has a large number of vowels in comparison to MSA/CA or any Arabic dialect variety. For example, while American English includes 16 vowels and each of SSBE and AusE have 20 vowels (details in the subsequent section), the SA variety has only eight. English is considered a qualitative language as the distinction between is based on vowel quality rather than quantity, with distinctions between vowels corresponding to differences in the tongue's position and height (Wells, 1982). With this in mind, Arabic speakers frequently encounter difficulties perceiving and articulating English vowels as vowel contrasts in their L1 are based more on quantity (vowel duration) than quality (Khattab, 2007). Moreover, English distinguishes between some vowel pairs (e.g., /iː ɪ/, /uː ʊ/) by utilising the characteristic of "tenseness". Mitleb (1984) argues that the duration difference between tense and lax vowels in English is attributed to an articulatory demand. This demand necessitates that the vocal organs maintain the configuration of tense vowels longer than lax vowels. The following provides a general description of three different varieties of English (SSBE, AusE, AmE).

2.6.2.1 SSBE vowels

SSBE has a more intricate vowel system in comparison to some varieties of English, primarily because of the absence of rhoticity (McMahon, 2002). It contains twenty vowels, including 12 monophthong vowels (/ɪ iː e æ ʌ ɜː ə ʊ uː ɑː ɔː ɒ/) and 8 diphthongs /ɪə eɪ eə aɪ əʊ ʊə ɔɪ aʊ/ (Reetz & Jongman, 2020; Wells, 1982). Almurashi (2022) conducted an acoustic analysis on 11 SSBE monophthongs (ɪ iː e ɑː ɔː ʊ uː æ ɒ ʌ) using both static and dynamic approaches. SSBE participants from both genders pronounced these vowels in monosyllabic and disyllabic syllables in the carrier phrase 'say ___ again'. There were apparent distinctions between tense and lax vowels. Figure 2.4 illustrates the normalised midpoint of F1 and F2 formant values of English vowels produced by SSBE speakers. There was a noticeable separation between most of the vowels, although there was some overlap in the vowel space of a few of them (e.g., /ɒ/, /ɑː/, /æ/, and /ʌ/). In addition, lax vowels exhibited greater centralisation than their longer counterparts. Regarding the additional cues (vowel duration, F0, and F3), the study showed that F0 played a more significant role than F3 and vowel duration.

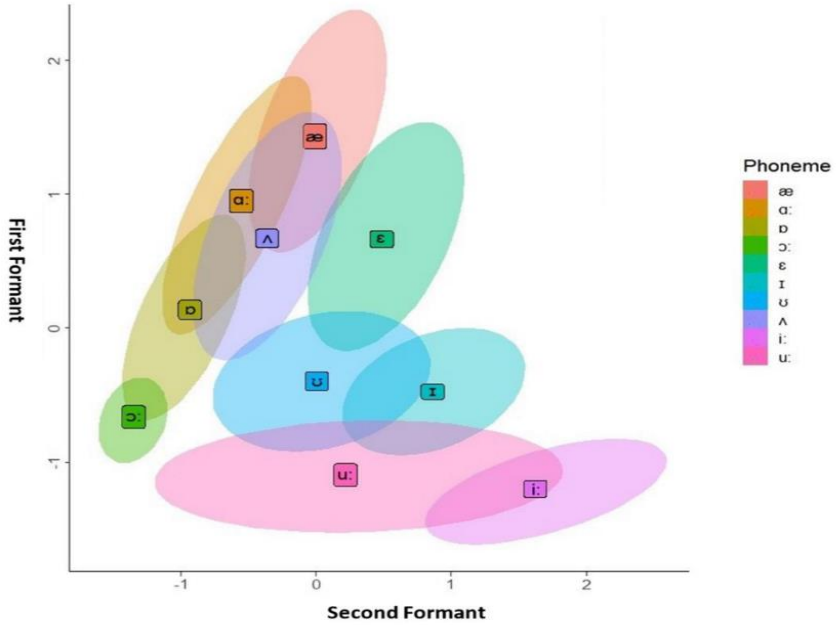


Figure 2.4 A scatter plot illustrating the normalised F1 and F2 midpoint values for 11 SSBE monophthongs produced by SSBE speakers (Almurashi, 2022).

The figure below displays the closing diphthongs /ɪə aɪ əʊ ɔɪ aʊ/ and the centring diphthongs /eɪ eə ʊə/ of SSBE in the vowel space. Given that SA does not have diphthongs in its vowel system, it is anticipated that the SA learners who participated in this study will associate SSBE diphthongs with the closest L1 monophthongs they already recognise, apart from /aɪ/ and /aʊ/.

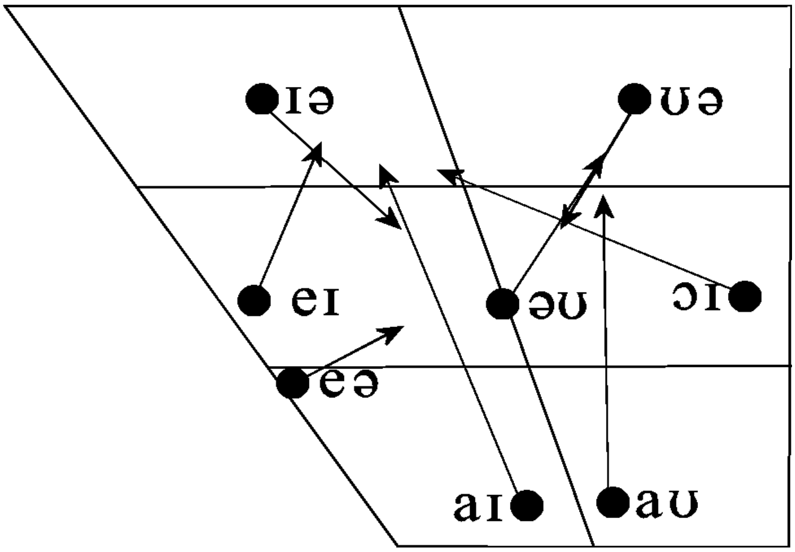
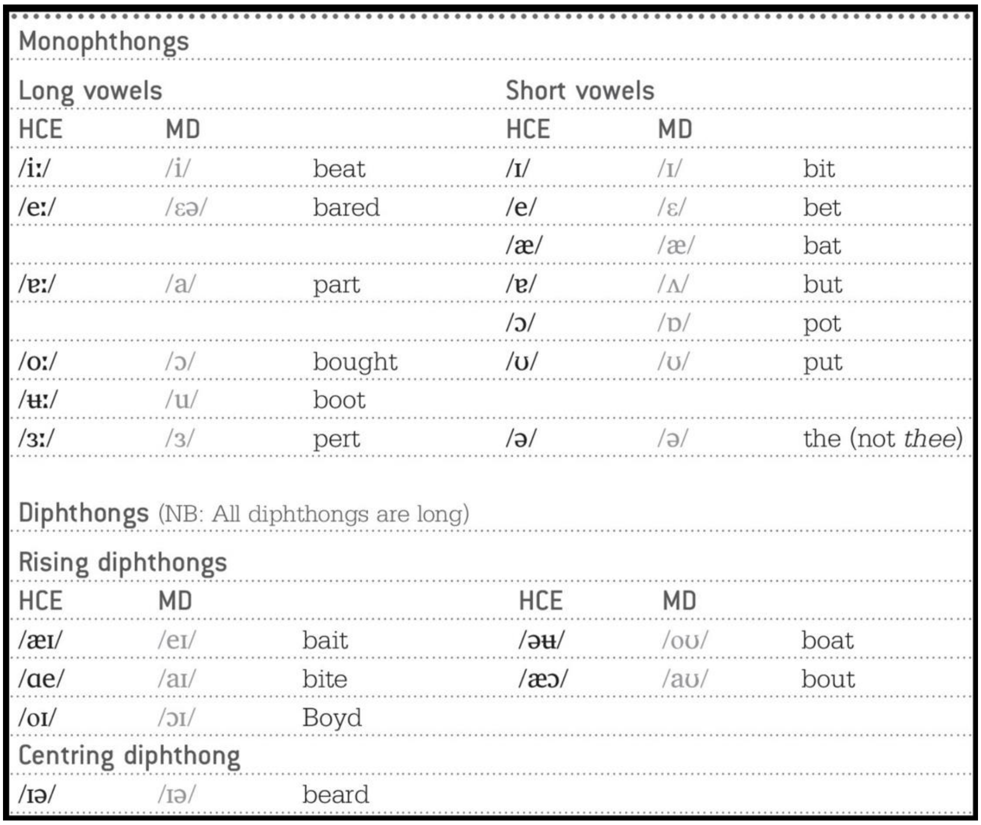


Figure 2.5 SSBE diphthongs, including closing and centring (Roach, 2010).

2.6.2.2 AusE vowels

Like SSBE, AusE has a more significant number of vowels (20) than other English varieties, which is attributed to its non-rhotic nature. Mitchell and Delbridge initially developed the phonemic vowel transcription system for AusE in the mid-20th century, drawing from south-eastern British English due to historical ties. However, many AusE vowels are not accurately represented by this system, known as the MD system. Harrington, Cox, and Evans later introduced an alternative (the HCE system) in 1997, specifically designed for AusE based on acoustic research. This system aligns more closely with the actual production of AusE vowels. Figure 2.6 displays the monophthongs and diphthongs of both systems. The updated phonemic vowel transcription system by (Harrington et al., 1997) is typically used by researchers since it is widely acknowledged for accurately representing Australian vowels.



Monophthongs					
Long vowels			Short vowels		
HCE	MD		HCE	MD	
/iː/	/i/	beat	/ɪ/	/ɪ/	bit
/eː/	/ɛə/	bared	/e/	/ɛ/	bet
			/æ/	/æ/	bat
/ɐː/	/a/	part	/ɐ/	/ʌ/	but
			/ɔ/	/ɒ/	pot
/oː/	/ɔ/	bought	/ʊ/	/ʊ/	put
/ʉː/	/u/	boot			
/ɜː/	/ɜ/	pert	/ə/	/ə/	the (not *thee*)
Diphthongs (NB: All diphthongs are long)					
Rising diphthongs					
HCE	MD		HCE	MD	
/æɪ/	/eɪ/	bait	/əʉ/	/oʊ/	boat
/ɑe/	/aɪ/	bite	/æɔ/	/aʊ/	bout
/oɪ/	/ɔɪ/	Boyd			
Centring diphthong					
/ɪə/	/ɪə/	beard			

Figure 2.6 AusE vowels that are included in the phonetic transcription systems of MD and HCE, as presented by Cox (2019).

The monophthongs of AusE include /ɪ e ə æ ɐ ɔ ʊ iː ɜː ɐː oː ʉː eː/ while its diphthongs are /æɪ ɑe oɪ əʉ æɔ ɪə ʊə/[30] . Certain vowel pairs in AusE are distinguished predominantly by length rather than spectral properties (Willoughby & Manns, 2019). (Cox, 1996) emphasised that the /ɐː, ɐ/ vowel distinction is founded on phonemic length, with the duration being the critical determinant. Similarly, other vowel contrasts like /eː, e/ and /ɪ, ɪə/ rely on duration as a primary identifier, especially in closed syllables, as seen in words such as "bid" vs "beard" (Cox, 2006). However, the distinction between the /iː, ɪ/ pair is not only predicted by length. Bernard (1970)

[30] Harrington et al. (1997) notes that the vowels /e:/ and /ɪə/ are in a similar position to /ɪ/ and /e/ in the vowel space and are often produced as long monophthongs or diphthongs with a centring glide, especially in closed syllables.

highlighted the presence of an onglide in /iː/ that is absent in /ɪ/, distinguishing between them based on both duration and spectral properties. Figure 2.7[31] depicts the impressionistic locations of the stressed AusE monophthongs within the vowel chart and some IPA cardinal reference vowels positioned at the space's edges (Cox & Fletcher, 2017). On the other hand, Figure 2.8 depicts the AusE monophthongs on the F1/F2 plane through an acoustic vowel space, as documented in (Burnham et al., 2011). This analysis collected information from 17 young female speakers in Sydney. The vowels were investigated within the /hVd/ reference context of the AusTalk corpus.

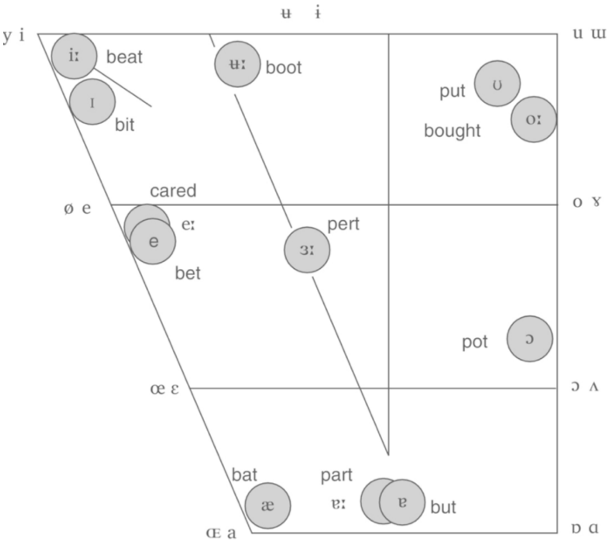


Figure 2.7 The AusE monophthongs in relation to the IPA cardinal chart (Cox & Fletcher, 2017)

[31] It is important to note that Figure 2.8 provides a schematic illustration of vowel positions, not an acoustic characterisation. The grey circles in the figure do not represent vowel variants; rather, they indicate the approximate location of the vowels (Cox & Fletcher, 2017).

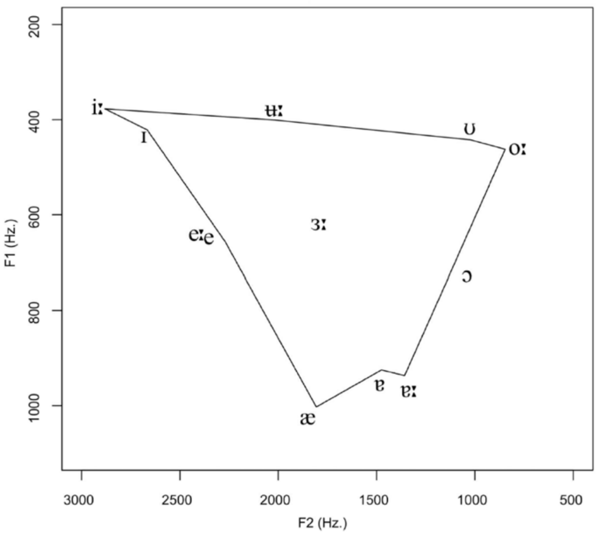


Figure 2.8 The F1/ F2 values for the AusE monophthongs (Burnham et al., 2011).

The diphthongs of AusE /æɪ ɑe oɪ əʉ æɔ ɪə ʊə/ have more dynamic articulatory movements than the monophthongs (Elvin et al., 2016). The vowel /əʉ/ has undergone notable shifts in AusE since the 1960s. According to Cox & Fletcher (2017), the vowel /əʉ/ shows regional differences in terms of the trajectory of the glide through the vowel space and the positions of its component sounds. The vowel /ʊə/ is uncommon in present-day AusE, so words like ‘tour’ and ‘pure’ often being produced as [tʉːə] and [pjʉːə] (Cox & Fletcher, 2017). Figure 2.9, derived from Cox et al. (2010) research, depicts the trajectories of AusE diphthongs from their onset to their end in relation to the IPA cardinal chart. However, according to Cox and Fletcher (2017), these straight lines may not accurately depict the prevalent movement of vowels within the vowel space.

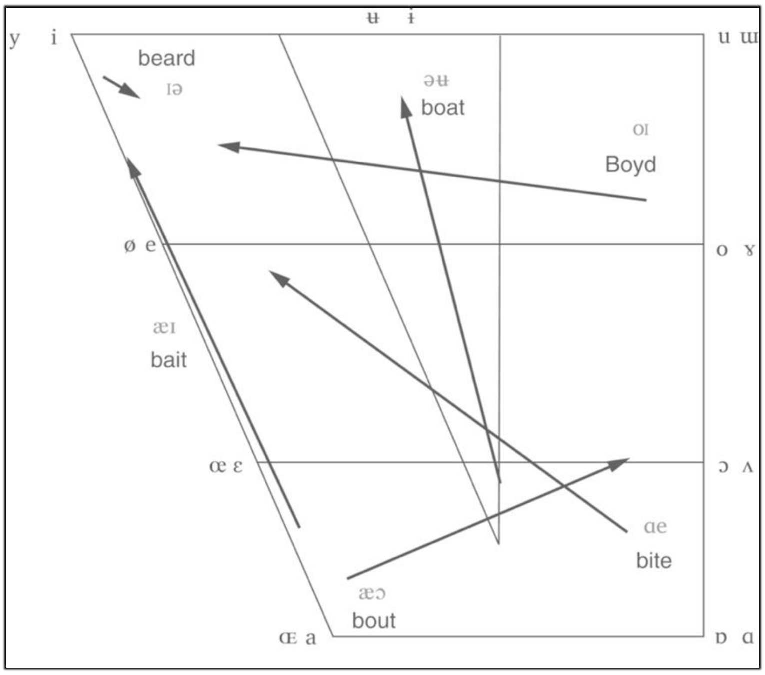


Figure 2.9 AusE diphthongs in relation to the IPA cardinal chart (Cox & Palethorpe, 2010b).

2.6.2.3 AmE vowels

The vowel system in AmE includes 11 monophthongs (/i ɪ ɛ æ ʌ ɝː ə ʊ u ɔ, ɑ/) and 5 diphthongs (/eɪ aɪ aʊ oʊ ɔɪ). Various publications use different symbols to represent the same AmE vowel sounds, resulting in inconsistencies in scholarly works. Yavas (2020) compared the symbols he used with those in other publications for both monophthongs and diphthongs. For clarity, the researcher associated each vowel symbol with a keyword. The table below represents these symbol variations:

Symbols used in Yava (2006, 2020) for AmE vowels	Keyword	Symbols used in other publications
/i/	beat	(/iː/, /ij/, /iy/)
/ɪ/	bit	
/e/	bait	(/eɪ/, /ej/, /ey/)
/ɛ/	bet	
/æ/	bat	
/ʌ/	bus	(/ə/ in unstressed syllables)
/ɑ/	pot	(/ɑː/)
/ɔ/	cloth	(/ɔː/)
/o/	boat	(/oʊ/, /ow/)
/ʊ/	book	
/u/	boot	(/uw/, /uː/)
/aɪ/	bite	(/aj/, /ay/, /ai/)
/aʊ/	bout	(/aw/, /au/, /ɑʊ/)
/ɔɪ/	voiced	(/oy/, /oj/, /ɔj/, /ɔy/, /oɪ/, /oi/)

Table 2.4 Symbols for American English vowels used in (Yavaş, 2020) and other publications.

According to Yavas (2006), AmE monophthongal vowels /i/ (*beat*) and /u/ (*boot*) have some level of diphthongisation despite being classified as "simple" vowels. Owing to this, some publications have represented these vowels using symbols like /ij/, /iy/, and /uw/. Moreover, the vowels /e/ (as in "bait") and /o/ (as in "boat") exhibit an increased level of diphthongisation, leading many references to adopt symbols such as /eɪ/, /ej/, /ey/ for /e/ and /oʊ/, /ow/ for /o/. Ladefoged and Disner (2012) argue that the vowels in 'bait'(/e/, /eɪ/) and 'boat' (/o/, /oʊ/) should be classified as diphthongs, despite their sound shifts being less distinct than those of the three primary diphthongs of AmE /aɪ/, /aʊ/, and /ɔɪ/. These three main diphthongs appear in any word position, and each begins with a stressed vowel and ends with a high vowel sound, as highlighted by Yavaş (2020). Unlike SSBE and AusE, AmE has a distinct rhotacisation feature. This relates to r- coloured vowels whose third formant frequency undergoes a reduction (Ladefoged & Maddieson, 1996). Notable r-coloured vowels in AmE are /ɝ/ ("bird"), /ɚ/ (unstressed "better"), /ɑr/ ("car"), and /ɔ˞/ ("north") (Clark et al., 2007).

The most frequently referenced study on the acoustics and perception of AmE vowels was conducted by Peterson and Barney (1952). In the phonetic context of /hVd/, they recorded two repetitions of ten vowels spoken by 33 male speakers, 28 female speakers, and 15 children. They measured each token's first three formants, formant amplitudes, and fundamental frequency. Hillenbrand et al. (1995) identified limitations in Peterson & Barney's (1952) study: it lacked spectral change patterns, duration measurements, and subject dialect screening, which could have affected acoustic outcomes. Addressing these issues, Hillenbrand et al. examined a larger, dialect-screened group of AmE speakers. Around 87% of the participants were from Michigan, with the rest from the inland northern regions. They measured vowel durations, F0 contours, and formant frequencies for 12 vowels in both the same phonetic context as Peterson and Barney (1952) and in isolation. Figure 2.10 represents the acoustic vowel charts developed by Peterson and Barney (1952) and Hillenbrand et al. (1995) for ten AmE vowels. It exemplifies the Northern Cities vowel shift found by Hillenbrand et al., where there is an upward movement in the vowel /æ/ and a change in the back vowels towards a more front and lower position, particularly /ɑ/ and /ɔ/, compared to the patterns portrayed by Peterson and Barney.

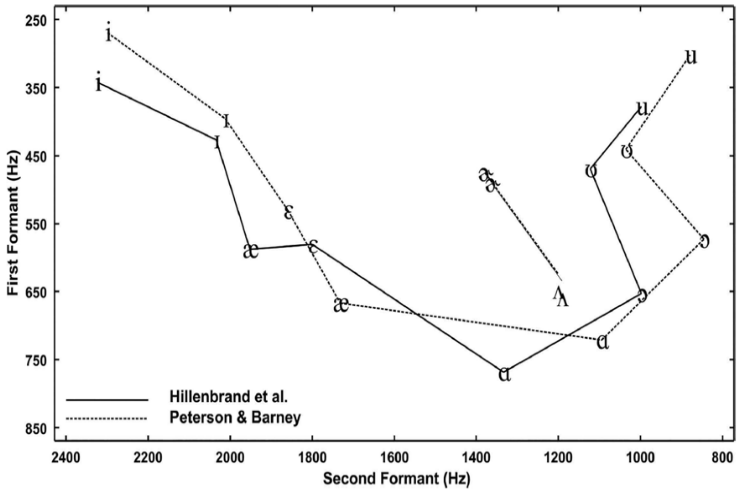


Figure 2.10 Acoustic vowel diagrams from Peterson and Barney (1952) and Hillenbrand et al. (1995).

Table 2.5 provides a summary of the vowels found in the English varieties previously reviewed (SSBE, AusE, AmE), along with relevant keywords.

Lexical Set	SSBE	AusE	AmE
FLEECE	/iː/	/iː/	/iː, i/
KIT	/ɪ/	/ɪ/	/ɪ/
TRAP	/æ/	/æ/	/æ/
DRESS	/e/	/e/	/ɛ/
NURSE	/ɜː/	/ɜː/	/ɝː/
lettER	/ə/	/ə/	/ə/
STRUT	/ʌ/	/ɐ/	/ʌ/
GOOSE	/uː/	/ʉː/	/u/
FOOT	/ʊ/	/ʊ/	/ʊ/
NORTH	/ɔː/	/oː/	/ɔ/
LOT	/ɒ/	/ɔ/	/ɑ/
START	/ɑː/	/ɐː/	/ɑɚ/
FACE	/eɪ/	/æɪ/	/eɪ/
PRICE	/aɪ/	/ɑe/	/aɪ/
CHOICE	/ɔɪ/	/oɪ/	/ɔɪ/
GOAT	/əʊ/	/əʉ/	/oʊ/
MOUTH	/aʊ/	/æɔ/	/aʊ/
NEAR	/ɪə/	/ɪə/	/ɪɚ/
SQUARE	/eə/	/eː/	/ɛɚ/
CURE	/ʊə/	/ʊə/	/ʊɚ/

Table 2.5 List of vowel sounds in SSBE, AusE, and AmE.

2.7 Arabic learners' perception and production of English vowels

Many earlier studies have shown the prevalent difficulties Arabic learners have with perceiving and producing English vowels (e.g., Alotaibi, 2018; Alshangiti, 2015; Almbark & Hellmuth, 2015; Almbark, 2012; Alghamdi, 1998; Abou Haidar, 1994; Munro, 1993). This difficulty is typically attributed to English having a denser vowel space than Arabic (Almurashi, 2022). Furthermore, Arabic is known for its emphasis on quantity, raising concerns about potential challenges when dealing with a qualitative language like English (Munro, 1993; Khattab & Al-Tamimi, 2008).Given that the difficulties encountered by Arabic learners in learning English vowels may vary by dialect (Newman & Verhoeven, 2002; Alotaibi, 2018), this section reviews only the studies that focus on SA learners' perception and production abilities of English vowels (without including any training in production or perception).

Alshangiti (2015) investigated the perception and production of English vowels and consonants (SSBE) by 26 SA learners from Riyadh and Jeddah who were classified as high proficiency (HP) and low proficiency (LP). The table below reports the tasks related to vowel[32] perception and production undertaken by the participants:

Tasks	Procedures
Identification test (In a quiet condition)	Participants heard recordings of vowels embedded in /hVd/ words and were tasked to identify each by selecting one of 17 test words. On the screen was a representation of all 17 vowels in the /hVd/ format, each coupled with a commonly associated word, such as "hide" likened to "bite," "hoed" likened to "code," and "haired" akin to "paid."
Identification test (In a noise condition) Identifying duration-equated vowels in noise	Each vowel was repeated twice at three distinct SNR levels: 0 dB, -5 dB, and -10 dB. The responses in these tests were collected using the same method as the vowel identification test administered in a calm environment.

[32] See Alshangiti (2015) for a review of the procedures, measures, and detailed results pertaining to the perception and production of consonants.

Production tests	Participants recorded each /hVd/ word three times using the carrier sentence "Say ___ again," while each word was displayed on a separate PowerPoint slide. They also read the passage "The north wind and the sun" (IPA Handbook, 1999).

Table 2.6 Tests used to evaluate the perception and production of English vowels by Saudi learners in Alshangiti (2015).

Using a vowel intelligibility test, 9 SSBE listeners identified the vowels produced by Arabic speakers based on a presentation mirrored the perceptual task and highlighted the 17 /hVd/ vowels and their rhyming words. For accent evaluation, SSBE listeners assessed an excerpt (the first sentence) of "The North Wind and the Sun" recorded by participants. On the 7-point Liberty scale, a score of 1 indicated a very native-like accent and a score of 7 showed a markedly non-native accent.

In terms of the results, all Arabic learners, regardless of their level of proficiency, faced more difficulty in identifying vowels than consonants in quiet and noisy conditions. Considering vowel identification in quiet, HP listeners registered a 68% score, while LP listeners reached 46% (when it came to consonants, HP listeners scored 82%, and LP listeners attained 72%). The table below displays the perception performance of each vowel (in quiet condition) by LP Saudi students[33]. These results are of great significance and will be compared to those of the current thesis, which also involves LP Saudi learners in quiet conditions.

[33] The emphasis here is on reporting the perceptual outcomes of LP learners. See Alshangiti (2015) for a review of the procedures, measures, and results pertaining to the perception and production of consonants.

SSBE vowels	Perception Performance
/ɑː/	85%
/æ/	79%
/iː/	74%
/e/	69%
/ɔː/	62%
/aʊ/	59%
/eɪ/	59%
/ɔɪ/	56%
/ʊ/	51%
/aɪ/	46%
/ɒ/	3%
/ɪ/	8%
/əʊ/	18%
/ɛə/	18%
/ʌ/	31%
/uː/	36%
/ɜː/	44%

Table 2.7 Results of the vowel identification task for LP learners

From the given data, the vowels /ɑː æ iː, e, ɔː, aʊ, eɪ, ɔɪ, ʊ/ are seemingly easier for learners to perceive (above 50%), whereas /ɒ, ɪ, əʊ, ɛə, ʌ, uː, ɜː aɪ/ (below 50%) pose more significant challenges. Considering the difficult vowels, the vowel /ɪ/ was frequently mistaken for /e/ in 72% of cases. The vowels /uː/ and /əʊ/ were identified as /ʊ/ in 54% and 36% of instances, respectively. The vowel /ɒ/ was predominantly perceived as /ɑː/ (38%) and as /ʌ/ (18%), while the vowel /ɛə/ was often confused with /ɜː/ (28%). Moreover, the vowel /ʌ/ was also misidentified with /ɒ/ (15%), /ɑː/ (21%), and /ɜː/ (21%), while the vowel /aɪ/ was confused with /ɪ/ (26%) and /eɪ/ (18%). Even though the vowels /e/ and /ʊ/ had identification rates of 69% and 53%, respectively (classifying them as "easy"), the high confusion rates with /ɪ/ for /e/ and /əʊ/ and /uː/ for /ʊ/ suggests that /ɪ/ and /ʊ/ should be regarded as challenging vowels for Saudi learners. With this in mind, the identification of vowels (easy, difficult) in Alshangiti's (2015) study was determined as follows:

- Easy vowels: /ɑː, æ, iː, ɔː, aʊ, eɪ, ɔɪ/
- Difficult vowels: /ɒ, ʌ, ɪ, e, əʊ, uː, ʊ, ɛə, ɜː, aɪ/

Aside from /ʌ/ and /aɪ/, it seems that the classification of vowels as simple or challenging in Alshangiti (2015) depends primarily on whether the L2 vowels are phonetically similar or dissimilar to Arabic vowels (/u a i uː aː iː eː oː/). For example, the SSBE vowels /ɔː aʊ eɪ ɔɪ/ are phonemically unshared vowels, as they do not have corresponding IPA symbols in Arabic. However, they are phonetically similar vowels, and the findings indicate that learners can easily perceive them. Conversely, the phonemic vowel /uː/ is phonemically shared, akin to the Arabic /uː/, yet phonetically unshared vowel and the findings indicate that learners encounter difficulty in perceiving this vowel. Refer to Table 2.11 for the comprehensive classification of shared and unshared vowels between Arabic and English, considering both phonemic and phonetic viewpoints.

With respect to Arabic learners' ability to differentiate English vowel contrasts, Alotaibi (2018) conducted a study on how Arabic speakers from two varied dialects, Saudi Arabic (SA) and Tunisian Arabic (TA)[34], discriminate the American English vowel pairs /æ/-/ʌ/, /i/-/ɪ/, and /u/-/ʊ/. The study mainly explored the extent to which variations in the learners' native dialects influenced their capacity to discern these English vowel contrasts. Both groups engaged in an AX discrimination task involving the L2 contrasts in four potential trial sequences (AB, BA, AA, BB). In general, the results indicated that discriminating English vowel contrasts varied across dialects, highlighting the need to control for dialect when establishing areas of difficulty in the perception and production of difficult English vowels.

Focusing on the reported findings for SA learners, the data revealed that the /æ/-/ʌ/ contrast was the most challenging to discriminate (with an error rate: of 82%), followed by /u/- /ʊ/ (with an error rate of 13%) and then /i/-/ɪ/ (with an error rate of 7%). These outcomes align with expectations since, even after undergoing perceptual training (using an identification task), Alshangiti (2015) observed that SA learners' discriminatory abilities after training remained unchanged for most of the SSBE vowel

[34] SA learners utilised the dialect spoken in Al-Majmaah, a city in the Riyadh region, while TA students utilised the dialect spoken in Chebika, a city in the Kairouan region.

contrasts[35]. Only few vowel pairs including /æ/-/ʌ/, /ɪ/-e/, /ɑː/-/ɒ/, and /ɑː/-/ʌ/ showed significant improvement.

Given the correlation between perception and production (e.g., Flege, 2003), it is expected that vowels identified as problematic in perception by SA learners would also present difficulties in production. In light of this, a review of Algethami's 2023 study on the production of SSBE monophthongs by SA learners is provided. Although Algethami's study centres on comparing the production patterns of SA learners with those of SSBE speakers, particularly in terms of achieving native-like production, a comprehensive evaluation of each vowel will assess its ease of articulation and possible challenges to production. This will be contrasted with Alshangiti's (2015) findings regarding the perception of easy and difficult vowels (in particular, the performance of LP learners).

Algethami (2023) examined the production of SSBE monophthongs (/ɪ iː ʊ uː ʌ æ ɑː ɒ ɔː e ɜː a ə/) by 16 Saudi male L2 learners in the United Kingdom. Participants were upper intermediate and were either enrolled in English universities or planned to enrol. Their speech was acoustically analysed, encompassing the measurement of vowel formant frequencies (at midpoint) and durations. These measurements were then compared with those of Arabic vowels produced by four Saudi Arabic speakers and English vowels produced by four SSBE speakers. L2 learners and English speakers were instructed to produce vowels in the /hVd/ context, which were incorporated into the carrier phrase (I say hVd again and again). Arabic speakers, on the other hand, produced the eight SA monophthongs (/iː i a aː u u eː oː/) in CVC contexts. The carrier sentence "ʕaqra cVc wadxul lilqaʔah" [I read cVc and enter the hall] was employed with the SA vowels. The results revealed that SA learners demonstrated proficient use of duration in their production of English monophthongs, except for the schwa sound, where they did not show the same shortening level as native speakers. Spectrally, Algethami showed that L2 speakers utilise a noticeably smaller spectral vowel space compared to SSBE speakers, demonstrating a closer

[35] Participants engaged in a category discrimination test, where they were presented with three /hvd/ words spoken by three different speakers. Among these, two words were the same while one word was different.

similarity to SA data. That is to say, L1 influence significantly affects the vowel production of SA learners. Primarily, they are not required to discern subtle differences when their vowel system is less dense and does not necessitate utilising the entire phonetic space.

Arabic learners did not exhibit apparent spectral differences in producing the English vowels /ʊ/ and /uː/ and produced /uː/ at a notably lower frequency than native English speakers. In contrast, the study revealed spectral differences between the vowels /iː/ and /ɪ/. In particular, the spectral values of the vowel sound /iː/ produced by learners fell within the range observed for both Arabic and English productions of the same vowel sound. The production of /ɪ/ overlapped with the spectral characteristics of SA /i/ and SSBE /ɪ/. These findings underscore the challenges of producing the vowels /ʊ/ and /uː/. Alshangiti (2015) also found these vowels challenging to identify for SA learners during the vowel identification test. In regard to /iː/ and /ɪ/, they share phonetic similarities with SA /iː/ and /i/. It appears that the former, /iː/, might be more easily produced by learners (even if it is not native-like) than the latter, /ɪ/. This could be attributed to variations in vowel length (long versus short) and tenseness (lax versus tense). Indeed, in terms of perception, Alshangiti (2015) observed that /iː/ (74%) was more readily identified than /ɪ/ (a mere 8%).

Additionally, Algethami found that SA learners produced the vowel /e/ with a notably higher F1 than the English group. Further, there was a noticeable overlap in their productions of /e/ and /ɪ/. Referring to the perceptual data presented by Alshangiti (2015), the vowel /ɪ/ was frequently confused with /e/ in 72% of the cases. This evidence strongly suggests that the vowels /e/ and /ɪ/ are problematic for learners both in perception and production. Concerning the /ɜː/ vowel, Algethami observed its production with distinct spectral features—which were more fronted and higher than by the English group. Furthermore, the articulation of /ɜː/ was closer to that of /e/ in the vowel spectrum. This underscores challenges in its production. Consistently, Alshangiti reported that learners had difficulty perceiving this vowel, with a 44% accuracy rate. These results demonstrate that Arabic speakers struggle with the vowels /e/ and /ɜː/ in both production and perception. Such difficulties presumably stem from the absence of comparable vowels in the learners' L1, at both phonemic and phonetic levels.

Regarding the vowels /ɔː/ and /ɒ/, these were significantly more fronted compared to those of the English speaker group, and learners' articulation of these vowels also showed greater fronting when compared to their closest SA vowel, /oː/. The realisation of /ɔː/ closely resembled the SA vowel /oː/ (see Figure 2.11). The overlapping areas between /ɔː/ and /oː/ in the diagram below illustrate this similarity. The /ɔː/ seems relatively easy for learners, even without a direct phonemic equivalent, due to its phonetic similarity to the Arabic vowel /oː/. The articulation of /ɒ/ slightly overlapped with Arabic /oː/, as shown in the figure below. This vowel should be problematic for SA learners because it does not correspond phonetically or phonetically to Arabic vowels. When contrasting these findings with perceptual performance, Alshangiti (2015) observed a high identification rate for /ɔː/ at 62%, whereas the identification for /ɒ/ was markedly low at merely 3%. From this, it is evident that /ɔː/ is relatively more straightforward for Arabic learners, whereas /ɒ/ presents challenges in perception and production.

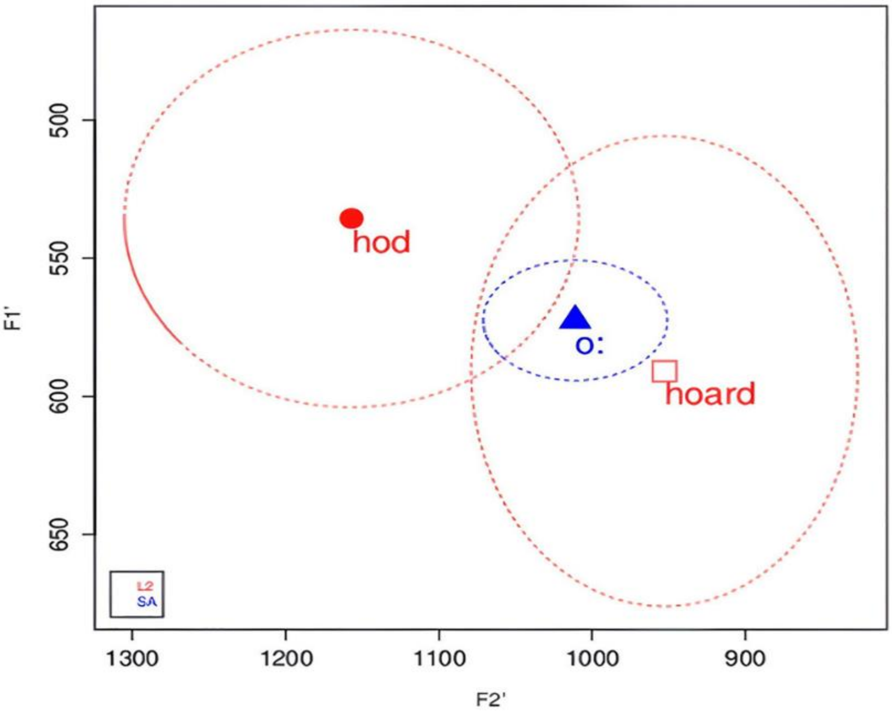


Figure 2.11 SA learners' production of /ɔː/ and /ɒ/ taken from Algethami (2023).

Additionally, there was an overlap between learners' production of English /æ/, /ɑː/, and /ʌ/ and their native vowel /aː/ (Figure 2.12). While there were spectral differences between the Arabic and English groups in /æ/ and /ɑː/ realisations (the Arabic group produced /æ/ higher and less fronted and /ɑː/ as more fronted than to the English group), there were no statistically significant spectral differences in the acoustic realisation of /ʌ/ between L2 and native speakers. In terms of perception, Alshangiti (2015) found the identification of /ɑː/ and /æ/ straightforward, with accuracy rates of 85% and 79% respectively. In contrast, the vowel /ʌ/ presented difficulties, achieving a score of 31%. These outcomes show that /ɑː/ and /æ/ are generally easy for SA learners in both perception and production. The discrepancy regarding /ʌ/, which is easier to produce according to Algethami (2023) but more challenging to perceive in Alshangiti (2015), could be attributed to variations in the participant groups of each study. However, considering its similarity to Arabic /a/ or /aː/, it can be anticipated that /ʌ/ is generally more achievable for SA learners.

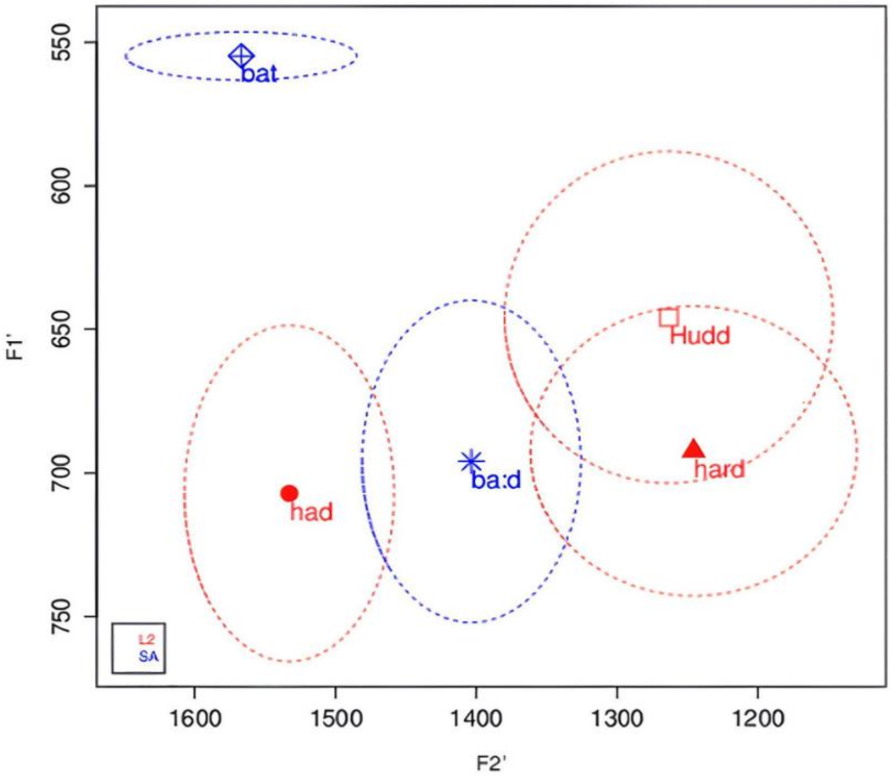


Figure 2.12 SA learners' production of /æ/, /ɑː/, and /ʌ/ taken from Algethami (2023).

Comparing the perceptual findings of Alshangiti (2015) with the production-oriented findings of Algethami (2023), it is evident that the phonetic similarities and differences between Arabic and English play a crucial role in determining the ease or difficulty with which Arabic learners produce and perceive English vowel sounds. Building on this, the current thesis explores whether the vowels deemed easy or difficult in the identification task stem from phonemic or phonetic disparities or both. The following table illustrates the shared and unshared vowels between Arabic and English used phonological and phonetic evidence from various studies on Saudi Arabic and English (Almurashi et al., 2022; Algethami, 2023; Alshangiti, 2015). This approach is taken since relaying solely on IPA symbols could obscure key differences. Furthermore, the focus of this study is not to attain native-like proficiency but rather intelligibility.

English vowels	**Phonological viewpoint**	**Phonetic viewpoint**
/iː/	shared (similar to Arabic /iː/)	shared (similar to Arabic /iː/)
/ɪ/	shared (similar to Arabic /i/)	unshared[36]
/uː/	shared (similar to Arabic /uː/)	unshared[37]
/ʊ/	shared (similar to Arabic /u/)	unshared[38]
/ɑː/	shared (similar to Arabic /aː/)	shared (similar to Arabic /aː/)
/æ/	shared (similar to Arabic /a/)	shared (similar to Arabic /a/)
/ʌ/	unshared	shared (similar to Arabic /a/ or /aː/)
/e/	unshared	unshared
/ɒ/	unshared	unshared
/ɔː/	unshared	shared (similar to Arabic /oː/)

[36] The English /ɪ/ typically does not phonetically match the Arabic's short /i/, but when a more relaxed variant of the Arabic sound is considered, the classification of /ɪ/ as a shared vowel is debatable (e.g., Khattab & Al-Tamimi, 2008; Almurashi et al., 2020).

[37] The SSBE vowel /uː/, similar to the Arabic /uː/, is a phonemically shared vowel. It is important to note, however, that the phonetic characteristics of the Arabic and English /uː/ differ substantially. See the production of Arabic and English /uː/ in Figure 2.13 taken from (Almurashi et al., 2023).

[38] The English /ʊ/ typically does not phonetically matched to Arabic's short /u/, but when a more relaxed variant of the Arabic sound is considered, the classification of /ʊ/ as a shared vowel is debatable (e.g., Khattab & Al-Tamimi 2008; Almurashi et al., 2020).

/ɜː/	unshared	unshared
/eə/	unshared	unshared
/əʊ/	unshared	unshared
/aɪ/	unshared	shared (similar to Arabic /iː/)
/aʊ/	unshared	shared (similar to Arabic /uː/)
/eɪ/	unshared	shared (similar to Arabic /eː/
/ɔɪ/	unshared	shared (similar to Arabic /oː/)

Table 2.8 The shared and unshared vowels between Arabic and English from both phonemic and phonetic perspectives.

The plot below, sourced from (Almurashi et al., 2023), provides a visual representation of the monophthongal vowels that are phonemically shared between HA and SSBE (/ɪ, i, ʊ, uː, æ, ɑː/) and those not shared (/ʌ ɒ ɔː ɛ/). It utilised the static model that measures formants (F1, F2) at the midpoint. The vowels were produced by HA, HA L2, and native English (NE) groups. It is worth mentioning that the phonetic production of HA /uː/ is quite distinct from NE /uː/, whereas HA /oː/ production aligns closely with the NE /ɔː/.

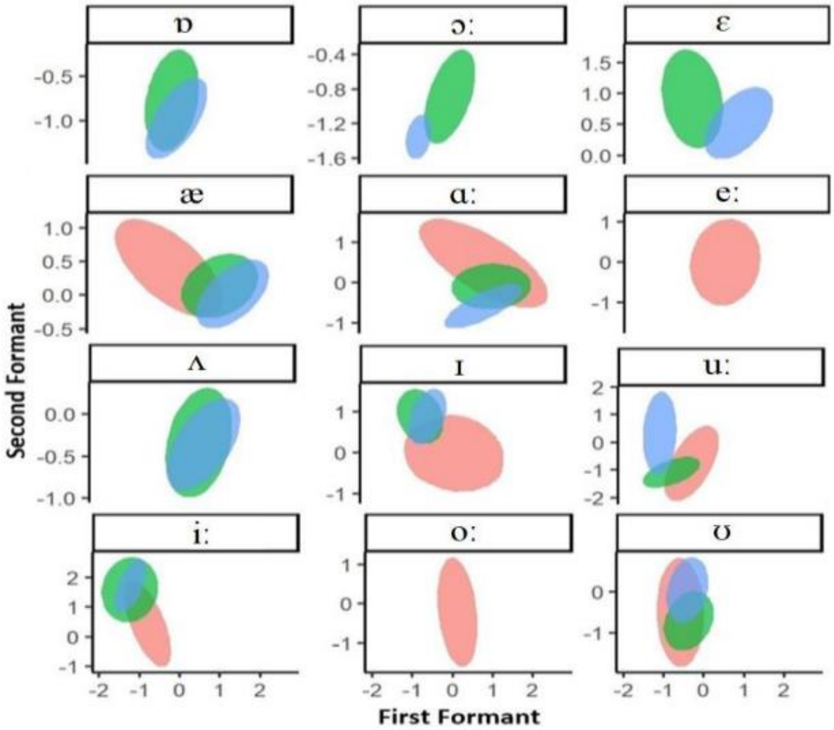


Figure 2.13 Normalised midpoints (LOBANOV Z-scores) of the first two formant values for vowels articulated by HA (colored in red), HA L2 (green), and NE (blue) participants, derived from (Almurashi et al., 2023).

Chapter 3. Methodology

This chapter opens with a summary of the study's design, followed by an overview of participants recruited for the study, the stimuli used for training and testing tasks, the training delivery mechanism, and the training plan. It then details the design of the training and evaluation tasks. Additionally, the chapter includes a brief summary of the ethical approval procedures and concludes with an exploration of the analysis methods implemented.

3.1 An overview of the study design

To evaluate the effectiveness and feasibility of HV training method in improving the perception of English vowels across diverse accents, participants were randomly allocated to three experimental groups: group A trained exclusively with the same L1 accent (SSBE), group B with multiple L1 accents (SSBE, AmE, AusE), and group C with a combination of L1 and L2 speakers (SSBE, AmE, SE). The training stimuli were identical across groups, except for the variations in accents.

The training consisted of 16 sessions distributed over a three-month period. Each session started with a production training phase consisting of 18 trials, where participants were instructed to record themselves producing a word and then listening to compare their pronunciation with that of a model speaker. This was followed by discrimination training, comprising 15 trials each for auditory and category discrimination tasks. Both discrimination tasks utilised a three-alternative forced choice format with feedback provided. Onscreen, three numbers appeared, and participants chose the number linked to the word that sounded different from the others in the auditory discrimination task, and the one with a distinctly different vowel phoneme in the category discrimination task. In the identification training, each session included 18 three-alternative forced choice trials. The computer displayed a set of minimal pairs (e.g., meat, mitt, met) alongside a single auditory token (e.g., mitt) which could not be replayed. Participants then selected the word they believed corresponded to the sound they heard and received feedback. In the LingLab vowel matching game (RSE, 2021), participants interacted with a setup of 12 cards. They were instructed to click on a card to hear an auditory stimulus and then rapidly find the corresponding word card. Correct matches turned the cards green, while incorrect ones turned them red. For a detailed description of each training task design and the feedback mechanisms, refer to section 3.6.5.

To ensure all participants were fully prepared for the training, they attended three preparation sessions that covered the differences between Saudi Arabic and English vowel systems (British, American, Australian) and provided explanations about the training and testing tasks. To assess the training effects, the experiment consisted of three testing phases: an initial pretest, a mid-test after the first eight training sessions, and a posttest following an additional eight sessions of the subsequent training program. All tests used the same SSBE accent stimuli and included three perceptual tasks without feedback: identification (35 trials), auditory discrimination (28 trials), and category discrimination (30 trials). To examine generalisation, participants underwent two generalisation tests: the first featured new speakers of familiar accents (SSBE, SE), while the second introduced speakers of new accents (Indian, Chinese). Similarly, these tests include the same perceptual tasks: Identification (36 trials), auditory discrimination (30 trials), and category discrimination (30 trials). Figure 3.1 below provides a simplified overview of the experimental design.

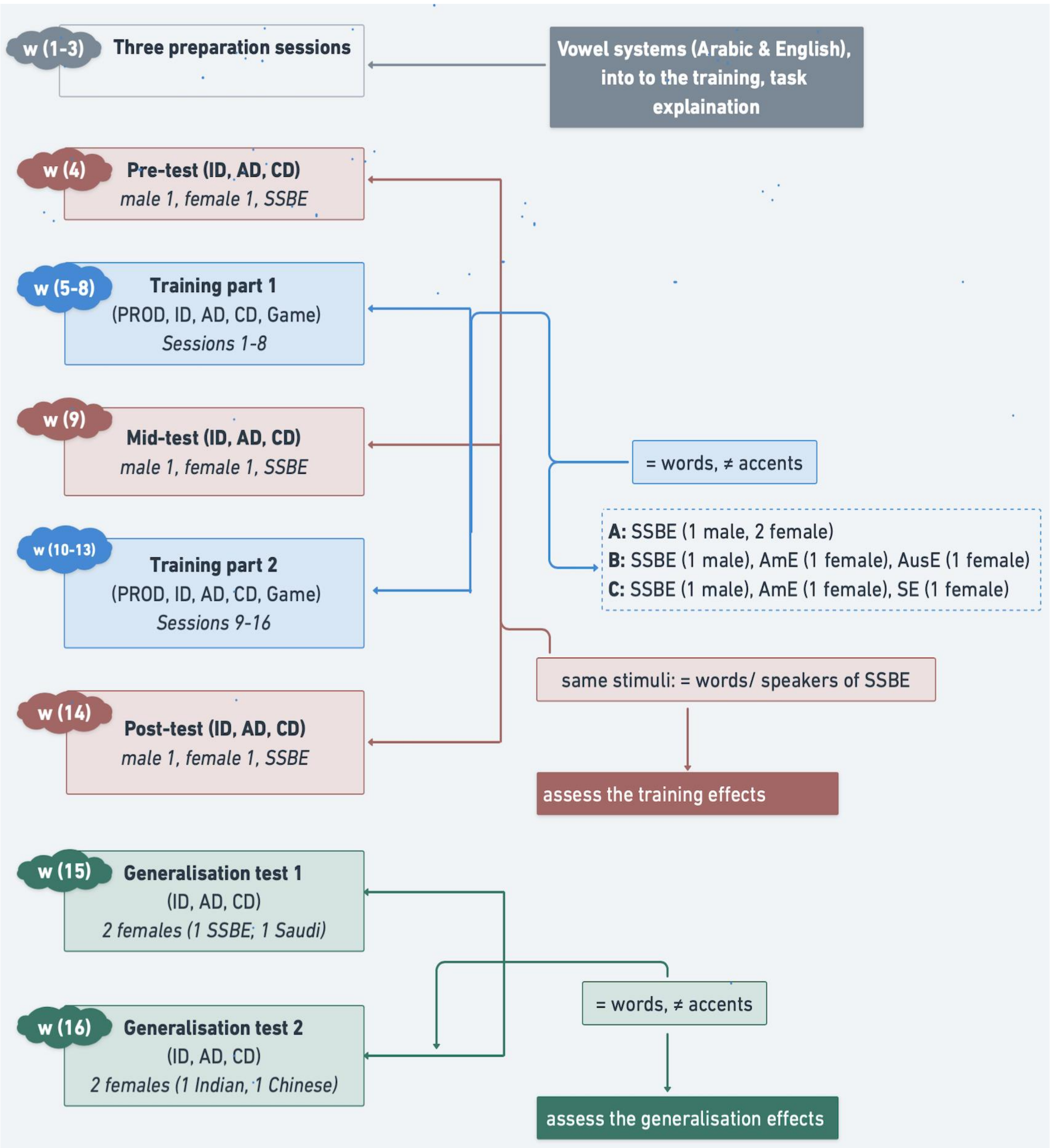


Figure 3.1 An overview of the experimental design

3.2 Population description and sampling

3.2.1 Recruiting L2 learners

Participants were female Arabic language learners of English between the ages of 20 and 25 from Alwajh College in the northwest region of Saudi Arabia. This university, like many others in Saudi Arabia, does not admit male learners due to cultural and religious restrictions. The participants had the same regional background and spoke the same northwest dialect. They were undergraduates studying seventh- and eighth-grade English Language and Computer Science levels. The training was considered an extracurricular course, with learners receiving credits for at least three subjects upon completion. The learners were awarded an e-certificate of accomplishment upon completion of the programme to boost their motivation. The study initially intended to conduct all sessions on campus and attracted a total of 180 participants. However, participation decreased significantly for a number of reasons. Some participants lost interest in conducting the training online due to COVID restrictions, others were affected by COVID, and some completed the training, but their data was excluded because, despite their assurances that they had completed the production task, their recordings did not show up on Labvanced along the training period. Ultimately, 126 participants completed the training online: 117 were at the beginner level, while 9 ranged from low intermediate to higher intermediate. The analysis concentrated exclusively on the performance of beginner learners. In fact, the study was originally aimed at intermediate learners, but most of the participants were found to be beginners. Concerns were raised regarding beginners' potentially insufficient lexical items. Despite these concerns, it was decided to proceed with training, incorporating weekly dictionary support. Additionally, the predominance of beginners sparked interest in investigating whether HV accent training could improve their perception of English vowels, even though this was not the central focus of the research. To answer the research questions presented in section 1.5, Arabic participants were randomly allocated to three experimental groups:

Group A	**38** beginner learners (A1) were exposed to a high variability paradigm with a single accent: they were trained on the SSBE accent.
Group B	**41** beginner learners (A1) were exposed to a high-variability paradigm with a variety of L1 Standard English accents in which they were trained on standard AmE, AusE, and SSBE.
Group C	**38** beginner learners (A1) were exposed to a high-variability paradigm featuring two standard L1 varieties and only one L2 variety during which they were trained on Saudi Arabic-accented English (SA-E), AmE, and SSBE.

Table 3.1 The distribution of participants across the three training conditions.

To guarantee that participants properly completed the training and testing sessions, comprehensive monitoring protocols were implemented across all sessions. Each session was systematically recorded on Labvanced, capturing both the start and end times. For each perceptual task—identification, auditory discrimination, and category discrimination—response times were rigorously tracked and documented in milliseconds for every trial. This meticulous tracking ensured that participants were fully engaged with the stimuli, thereby validating their active participation. In the production training, participants were directed to produce each word before hearing it from the model speaker and then compare their productions to the model's. However, a major limitation of this task is that Labvanced only records the final production and does not document the steps leading up to it, making it impossible to verify that participants strictly followed the instructions. Furthermore, participants whose recordings were not visible on Labvanced were advised to switch to alternative devices. If the issue persisted, their data was excluded to preserve the integrity of the dataset, although, as mentioned earlier, they were allowed to complete the training.

3.2.2 Participant recruitment for recording training and testing stimuli

Three SSBE speakers were recruited for group A paradigm: one speaker from each of SSBE, AmE and AusE for group B and one from each of SSBE, AmE and SE for group C. The AmE speaker was shared between groups B and C and one of the SSBE speakers from group A was also used in both group B and C. This arrangement resulted in a total of six speakers recording the training stimuli, with each group featuring two female and one male speaker.

The decision to employ the SSBE in group A stemmed from its consistent application in prior research (e.g., Iverson et al., 2023; Lengeris & Hazan, 2010; Iverson et al., 2005), enabling meaningful comparisons. In contrast, group B was exposed to a range of L1 standardised accents— SSBE, AmE, and AusE— without preference. The use of these standardised accents does not indicate a favouritism over regional L1 accents. Instead, they are employed initially to increase exposure, thereby allowing for the potential incorporation of regional L1 accents in future research. Additionally, incorporating SE, SSBE and AmE in group C covered both L1 and L2 varieties, aligning closely with the thesis's objective of offering comprehensive accent exposure without bias. The SE accent was selected as the non-L1 accent for group C (rather than another L2-accented variety) because it acted as a familiar accent in the gen1 test. This choice ensures that the SE accent is familiar not only to group C but also to groups A and B, even though the latter groups did not hear it during training. Since all participants share the same L1 as the SE speaker, the SE accent is equally recognisable to all groups. Therefore, using this L2 variety ensures fairness and provides a consistent basis for assessing generalisation to a familiar accent across all participants.

For the testing stimuli, two SSBE speakers (one male, one female) recorded the pre-, mid-, and post-tests, with identical stimuli across all phases to measure training effects. These speakers were not heard during the training or generalisation phases. This decision is well-founded given that SSBE is the common accent familiar to all three training groups, despite their differences. Using SSBE in the tests allow for a fair comparison of results among the groups, ensuing that differences in outcomes are not influenced by unfamiliar accents, but rather by the training itself.

During the two generalisation tests (gen1 and gen2), all speakers were new to the participants, regardless of accent familiarity. The gen1 featured novel speakers with familiar accents, while the gen2 involved novel speakers with unfamiliar accents. The purpose of this is to assess the adaptability of the training to different accents, providing insight into how training extends to various accentual variations. This method diverges from that used in earlier HV training studies (e.g., Iverson et al., 2023; Lively et al., 1994) where typically, generalisation tests would involve one familiar speaker (having been heard during the training phase) and a new speaker, both sharing the same accent.

Specifically, the gen1 test comprised a new SSBE speaker (whose accent was heard in groups A, B, and C) and a new Saudi speaker (whose accent was only heard in group C). As explained, even though groups B and C were not exposed to Saudi Arabic-accented English, they should still be familiar with the Saudi speaker's recording of gen1 since they share the same L1 with this speaker. The two L2 Saudi speakers, whether the one who provided the stimuli for group C or the one who recorded the gen1 test, were competent L2 English speakers. According to Matsuura (2007) and Eisenstein & Berkowitz (1981), whatever technique is used to assess intelligibility, it is predicted that high-proficiency L2 speakers may provide more reliable intelligibility test data. The Saudi L2 speaker who recorded the training stimuli fulfilled the International English Language Testing System (IELTS) 8 intelligibility criteria on the speaking component[40]. The highest grading scale, band 8 in the IELTS test, is given to pronunciation when the L1 has a marginal effect on intelligibility (IELTS, 2022). Similarly, the Saudi L2 speaker who recorded the test stimuli is highly skilled and achieved a band 7 score. Based on the Speaking Band Descriptors of IELTS, this

[40] See IELTS Speaking Band Descriptors at https://s3.eu-west-2.amazonaws.com/ielts-web-static/production/Guides/ielts-speaking-band-descriptors.pdf

band indicates that the pronunciation is generally clear and easily understandable without much effort. It also implies that this speaker might share some positive pronunciation qualities associated with candidates who attain band 8.

On the other hand, the gen2 test included Chinese and Indian accents[41] that were unfamiliar to all the groups despite their training. This approach supports the notion that L2 learners benefit from exposure to a range of English varieties rather than being restricted to a standard variety like SSBE or AmE. In particular, it supports Kachru's (1992) *Three Circle Model*, detailed in section 111, classifies English speakers into three groups: *the Inner Circle, the Outer Circle,* and *the Expanding Circle*. The decision to use an Indian accent was made to assess learners' perception of English as spoken in countries within *the Outer Circle*, while the Chinese accent was intended to evaluate their perception of English in *Expanding Circle countries*. Although Kachru's (1986) tripartite approach has faced criticisms regarding the alignment of English speakers with the *Three Circle Model,* employing competent L2 Indian and Chinese speakers is likely to heighten learners' awareness of the wide range of English varieties. The table below provides a comprehensive summary of the speakers and accents used in the training and testing phases of the study.

[41] Like L2 Saudi speakers, Indian and Chinese speakers also demonstrated competency as L2 English speakers. The Indian speaker was bilingual, while the Chinese speaker achieved an IELTS score of 7.5 band.

	Group A	Group B	Group C
Pre-test:	male 1, female 1, SSBE	male 1, female 1, SSBE	male 1, female 1, SSBE
Training part 1	male 2, SSBE female 2, SSBE female 3, SSBE	male 2, SSBE female 3, AmE female 4, AusE	male 2, SSBE female 3, AmE female 5, SA
Mid-test:	male 1, female 1, SSBE	male 1, female 1, SSBE	male 1, female 1, SSBE
Training part 2	Same speakers a Training part 1	Same speakers a Training part 1	Same speakers a Training part 1
Post-test	male 1, female 1, SSBE	male 1, female 1, SSBE	male 1, female 1, SSBE
Generalisation test I	male 3, SSBE female 6, Saudi	male 3, SSBE female 6, Saudi	male 3, SSBE female 6, Saudi
Generalisation test II	female 7, Indian female 8, Chinese	female 7, Indian female 8, Chinese	female 7, Indian female 8, Chinese

Table 3.2 Speakers and accents used in the training and testing phases.

3.3 Stimuli

The stimuli recordings were conducted using Zencastr, a high-quality web-based podcast recording software. Participants were sent individual email invitations for a Zencastr session scheduled at their convenience. To ensure optimal clarity of voice recordings, all sessions were conducted in audio-only mode, although video was utilised to enable real-time communication between the researcher and the participants. Participants were required to use headphones and select a quiet environment to minimise background noise. They were presented with a PowerPoint that sequentially displayed each word on separate slides in a random order, which helped prevent list intonation and preserved natural speech patterns. The list primarily consisted of familiar words, with a few nonsense words, each paired with two well-known words (e.g., "lurd (birth, dirt)" and "coal (note, wrote)") to facilitate correct pronunciation. Recording sessions lasted approximately 20 to 35 minutes, depending on each participant's speed.

Upon completion, a WAV file with a 48 kHz sample rate was extracted for each speaker. Each speaker's TextGrid and sound file were simultaneously opened using Praat (Boersma & Weenink, 2021). The stimuli were then orthographically transcribed, and boundaries were added. To normalise the amplitude of the stimuli and ensure comparable loudness, three scripts were run: the first extracted the target words and added 100 milliseconds of buffer around each one, the second scaled all recordings to 65 dB, and the third smoothed the edges (Al-Tamimi, 2022). It is important to note that the stimuli were organised in a nested design, indicating that each speaker produced a specific stimulus.

The study used a large set of English vowels (seventeen vowels) as stimuli: /iː/, /ɪ/, /e/, /ʌ/, /æ/, /ɒ/, /ʊ/, /ɑː/, /uː/, /ɜː/, /ɔː/, /ɛə/, /eɪ/, /aɪ/, /aʊ/, /əʊ/ and /ɔɪ/. This is because, at least for vowels, training on a large set is often reported to be more helpful for adult learners than on a small dataset (Nishi & Kewley-Port, 2007; Iverson & Evans, 2009; Iverson et al., 2012). Additionally, Thomson (2018) and Uchihara et al. (2021) showed that training in two-way contrasts (e.g., 'sit' vs ‘seat') can be ineffective, particularly for vowels, because learner confusion patterns are seldom binary. It is preferable to train on a broader range of sounds (e.g., 'sit', 'seat', 'set',' sat', 'suit',' sought', and so on). Given this, it is reasonable to assume that Arabic learners would gain from exposure to a broad vowel set.

The primary criterion for selecting the training and testing stimuli was to include as many real words as possible, ensuring that they resembled natural conversational language closely. They were quite diverse, with vowels occurring in a variety of consonantal environments. There were few pairings available that contrasted the vowels (/ʊ/, /uː/, /ɛə/, /ɜː/, /ɔɪ/). Thereby, few nonsensical words were incorporated into perceptual tasks during both the training and testing sessions (refer to Appendices D and E). Given that participants in this study are beginners, it is reasonable to expect that they will be unfamiliar with even familiar/frequent words. Consequently, using a limited number of nonsensical words in this context is both logical and justifiable. Additionally, the nonsense words used were orthographically transparent. For example, 'u' represented /ʌ/ as in 'sut', 'gud', and 'shud'; 'air' represented /ɛə/ as in 'bair', 'bairt', 'tairn', and 'caird'; and 'oi' represented /ɔɪ/ as in 'doit' and 'toit'.

3.3.1 Training stimuli

The training stimuli were identical across all three training groups (A, B, and C) except for the speakers heard in each group (see section 3.1.2 for sufficient details regarding the speakers recruited to record the training stimuli).

3.3.1.1 Stimuli for identification and production tasks

The stimuli used in the production task were predominantly reflected in the identification task[43]. This approach allowed participants to familiarise themselves with the speaker models during the production phase before proceeding to the identification task. There was a total of 12 vowel sets (clusters) derived from the study's seventeen target vowels, with each set consisting of three vowels. This decision was made not only by considering the production and perceptual confusion patterns observed in previous research conducted on SA learners by Alshangiti (2015)[44] but also by allowing for a more extensive examination of a larger number of vowels (See Table 3.3 below). A concerted effort was made to include as many potentially confusing patterns as possible. Even while not every member of each set was found confused with every other member, the sets offered here enabled Arabic learners to cover as many of the production and perception confusions as possible. Exposing learners to a wide range of examples in these sets can help them develop a more comprehensive understanding of vowel production and perception. This will significantly enhance their knowledge in mastering vowels throughout the training sessions. Vowel sets 11 and 12 specifically were introduced to better handle rhotic accents like AmE and L2 English varieties, including Saudi English, Indian and Chinese.

[43] Keep in mind that the production data collected in the study were not analysed owing to time constraints.

[44] The following vowel clusters (sets) were identified as presenting challenges in both perception and production for SA learners:

* /iː/, /ɪ/, /e/, /æ/, /aɪ/, /eɪ/
*/æ/, /ʌ/, /ɑː/, /ɒ/, /aɪ/, /aʊ/
*/ɑː/, /ɒ/, /ʌ/, /ɔː/, /əʊ/, /ɜː/
*/uː/, /ʊ/, /aʊ/, /əʊ/, /ɜː/

	Combinations of confused vowel sets
1	/iː/, /ɪ/, /e/
2	/e/, /eɪ/, /aɪ/
3	/ʌ/, /æ/, /ɑː/
4	/ʌ/, /ɒ/, /ʊ/
5	/ɜː/, /ɑː/, /ɔː/
6	/ɜː/, /ɛə/, /e/
7	/uː/, /aʊ/, /əʊ/
8	/uː/, /ʊ/, /ʌ/
9	/aʊ/, /uː/, /ʊ/
10	/aʊ/, /ɔɪ/, /əʊ/
11	/ɜː/, /ɪə/, /ɛə/
12	/ɔː/, /ɑː/, /əʊ/[45]

Table 3.3 Vowel sets used in the identification and production training.

The 12 vowel sets listed above in Table 3.2 were distributed across the 16 training sessions (see Figure 3.2 below). During the initial 6 sessions, participants encountered two distinct vowel sets per session, each containing three vowels, totalling 6 vowels. Each set was produced by three different speakers, resulting in 18 minimal pair words per session (6 vowels x 3 speakers). For example, the first session included the sets (/iː/, /ɪ/, /e/) and (/e/, /eɪ/, /aɪ/), with each vowel having three different minimal pairs (e.g., /iː/: 'feel,' 'meat,' 'heed'; /ɪ/: 'mitt,' 'hid,' 'fill'), each produced by three different speakers (speakers 1, 2, and 3). Sessions 7 and 8 were designated as review sessions before the mid-test. Session 7 focused on the first 6 vowel sets, while session 8 covered the remaining 6 sets. Each set featured one minimal pair for each vowel, produced by a specific speaker (1, 2, or 3), resulting in a total of 18 minimal pairs per review session (3 words x 6 sets = 18 words). For instance, the set /uː/, /ʊ/, /ʌ/ included ‘who’d,’ ‘put,’ and ‘cud,’ produced by speakers 2, 3, and 1 respectively. Sessions 9 to

[45] The vowel /əʊ/ was mistakenly used in place of the /ʊə/ vowel.

14 revisited the same vowel sets heard in sessions 1 to 6, with new minimal pair words introduced for most vowels. Finally, sessions 15 and 16 were designated review sessions before preceding the post-test. presented in random order.

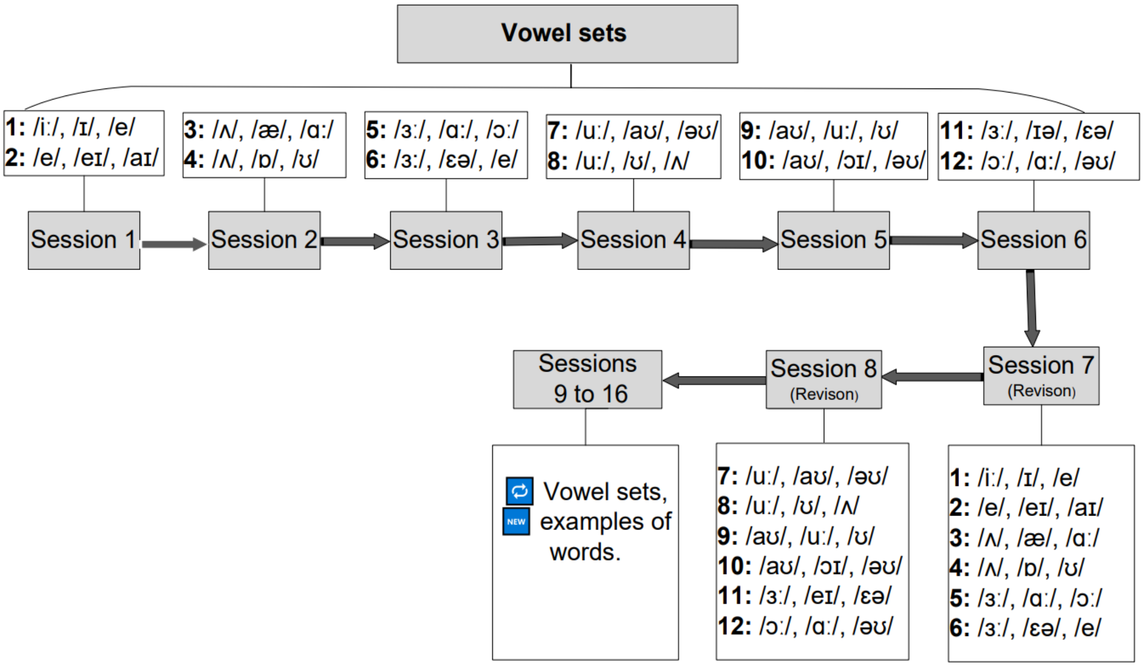


Figure 3.2 The distribution of vowel sets across the training sessions for identification and production tasks.

3.3.1.2 Stimuli for auditory and category discrimination tasks

There were 27 vowel pairs for the auditory and category discrimination training, as detailed in Table 3.4 below. These pairs were selected from the vowel sets identified as commonly confusing. Counterbalancing the order of presentation was employed to mitigate potential order. This involved systematically varying the order of vowel pair presentations across different trials. The target vowel was equally distributed in different positions within the sequence (the first, middle, last) across trials. Additionally, the order of vowel pairs was randomised to ensure that no specific sequence influenced the overall results.

	Vowel pairs
1	/iː/–/e/
2	/ɪ/–/e/
3	/iː/–/ɪ/
4	/e/–/eɪ/
5	/aɪ/–/eɪ/
6	/æ/–/ɑː/
7	/ʌ/–/æ/
8	/ʌ/–/ɒ/
9	/ɒ/–/ʊ/
10	/ʌ/–/ʊ/
11	/ɜː/–/ɑː/
12	/ɑː/–/ɔː/
13	/ɜː/–/ɛə/
14	/ɛə/–/e/
15	/ɜː/–/e/
16	/aʊ/–/əʊ/
17	/uː/–/ʌ/
18	/aʊ/–/ɔɪ/
19	/ɔɪ/–/əʊ/
20	/aʊ/–/uː/
21	/uː/–/ʊ/
22	/aʊ/–/ʊ/
23	/ɜː/–/eɪ/
24	/eɪ/–/ɛə/
25	/ɔː/–/ɑː/
26	/ɑː/–/əʊ/
27	/ɔː /–/əʊ/

Table 3.4 Vowel pairs used in auditory and category discrimination training

The vowel pairs listed above were distributed across the 16 training sessions, as shown in Figure 3.3. The initial six sessions consisted of 30 vowel pairs, with each session presenting 5 pairs. Each pair was produced by three different speakers, resulting in a total of 15 words per session (5 pairs x 3 speakers). The study featured only 27 vowel pairs; however, to achieve a balanced distribution of trials throughout the training sessions, three pairs (/ʌ/–/ʊ/, /aʊ/–/uː/, and /uː/–/ʊ/ were repeated. This provided a total of 30 vowel pairs for the initial sessions (1-6) and the later sessions (9 -14), with each sessions featuring 5 pairs. Session 7 and 8 served as revision periods, each including 15 vowel pairs. From sessions 9 to 14, the same vowel pairs as in the first six sessions were used, but with new words introduced for most vowels. Sessions 15 and 16 were treated as revision sessions.

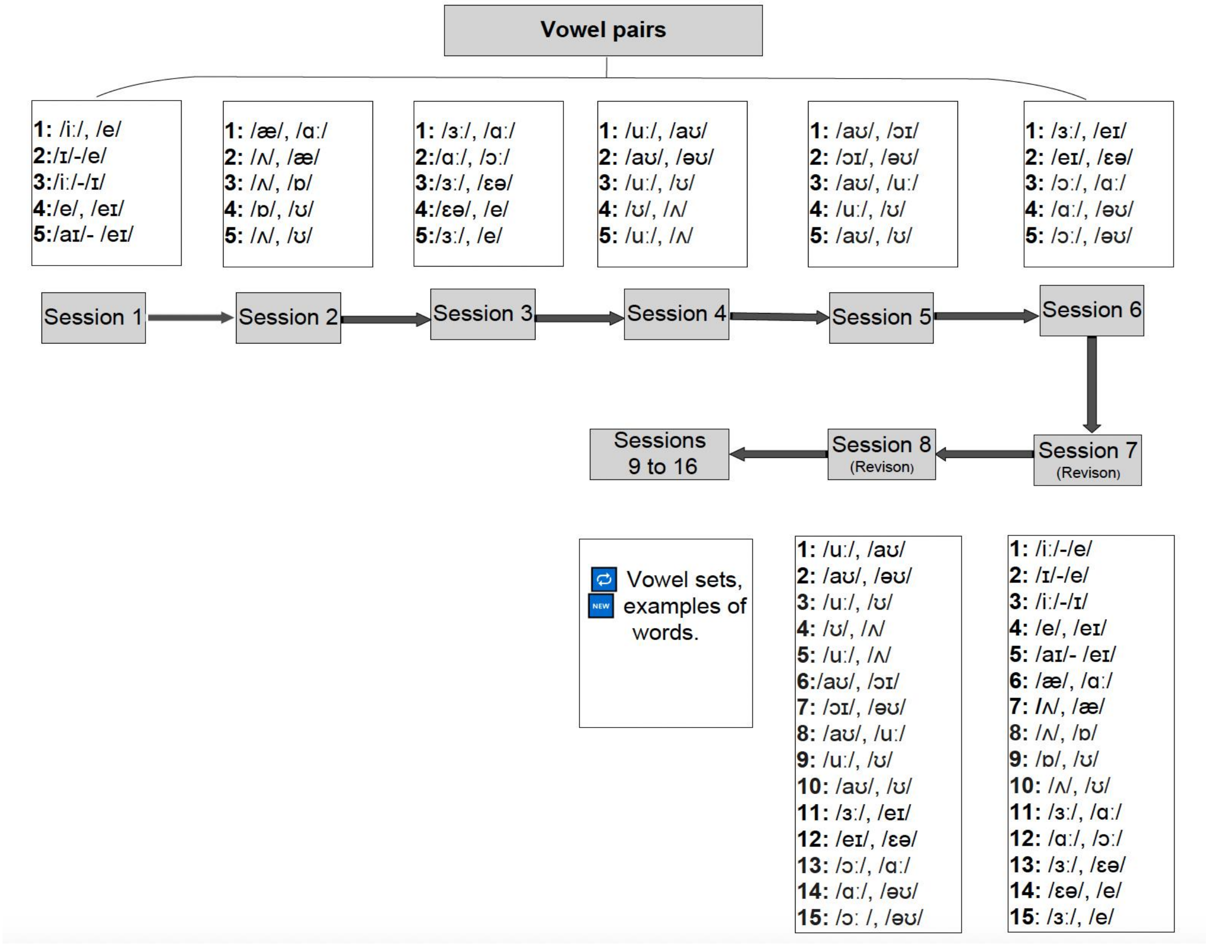


Figure 3.3 The distribution of vowel pairs across the training sessions for both auditory and category discrimination tasks.

3.3.2 Testing stimuli

The testing stimuli for the pre-, mid-, post-, and the generalisation tests consisted of the same sets and pairs of vowels that participants had experienced during the training. While the 12 vowel sets and 27 vowel pairs were spread out throughout the training sessions, each test session included all of them. The recordings for the pre-, mid-, and post-tests featured stimuli that were neither used in the training sessions nor the in the generalisation tests, to effectively assess the impact of the training. On the other hand, the generalisation tests incorporated new stimuli to evaluate the participants’ ability to apply what they learned to unfamiliar contexts.

The pre-test, mid-test, and post-test stimuli were identical (all recorded by SSBE speakers) so that participants' perceptual performance could be compared before, during, and after the training. The stimuli for the first and second generalisation tests were identical but produced by speakers with different accents. All testing stimuli, including the speakers heard by the participants, were identical across all training groups. To view test testing stimuli, see Appendix E in which

3.4 Training delivery modes

The initial plan was to conduct all training sessions on campus, but owing to COVID constraints, all training materials had to be prepared and delivered online. The desire of most learners to finish training sessions on weekends at their convenience was a further justification for offering training online. With the exception of an initial three-week preparatory session, which was conducted on campus, all training and testing sessions were conducted online.

Participants were assigned randomly to one of the three WhatsApp groups (A, B, C) that had been created. They were informed that they would receive identical training, but that each group would be listening to a different set of English speakers. None of the participants was aware of the accents employed in each training group.

The purpose of the WhatsApp group was to respond to participants' inquiries. Links to the training and testing sessions were not shared in WhatsApp groups.

Three WhatsApp *broadcast* groups (A, B, C) were created for efficient communication. This structure was designed to ensure that every participant directly received the relevant links for their assigned training and testing sessions. Participants remained within their designated groups: sharing training links with other participants was prohibited. Three Telegram channels were also established (A, B, C) to provide participants with written materials about the training and to keep them updated: the instructor posted each completed training or testing session. Furthermore, these channels included recorded videos created by the researcher, which provided detailed explanations of how to complete the training and testing tasks. While the researcher arranged initial campus meetings to explain these tasks to all participating students, the availability of recorded materials was deemed necessary as a reference for the students. This allowed them to revisit the explanations as needed, ensuring a clear understanding of the tasks.

3.5 Training plan

Before meeting with participants, the researcher was present on the Alwajh College campus for five days to meet with instructors, familiarise herself with the participants and arouse their interest in the training. To promote and advertise the training, the researcher created a poster using Canva, a free online graphic design platform (https://www.canva.com). The poster was created in Arabic (Appendix C1) and the translated English version can be found in Appendix C2. The poster included the aim of the training, the duration and place of the training, the target group, the training plan as well as the contact information. Registering for the training was made easily accessible using a QR code at the bottom of the poster.

The first three preparation sessions were held on campus. In the initial session, which took place during the first week, the three groups (A, B, C) were provided with a comprehensive introduction to the study they would be participating in. Participants were also given an overview of the distinction between sounds and letters and a description of vowel characteristics based on height, location, and rounding. Additionally, they were made aware of the significance of improving their production and perception of English vowels and that the number of vowels varies across different languages and accents. The second session was also conducted during the first week to provide participants with a comprehensive overview of the training and testing tasks[46]. It is crucial to ensure that participants clearly understand each task before they begin. This is essential for promoting active engagement in the training, optimising the learning process, and ultimately achieving the training's objectives. It was clear to learners that the figures given for task explanations were not representations of the actual design but provided a general overview of the training tasks. This is because explaining the tasks to participants happened before the training task design was completely finalised. In addition, in this preparation session, participants received guidance on choosing a username for the training and activating auto-play and microphone settings on their browsers. Two Zoom meetings were arranged to assist participants who were absent from the class meeting and experienced technical issues, including problems with enabling auto-play or managing the microphone on their devices

During the third preparation session (week 2), participants completed a questionnaire regarding their language history (see Appendix B) for the researcher to gain more information about their language skills (e.g., the amount of exposure to English and age of learning). The questionnaire was published as a Word document to the WhatsApp broadcast lists (A, B, and C) so all learners could access it. Those who could not edit the document on their smartphones or did not have a laptop or iPad in the class were given paper copies. They were required to submit a digital version of the questionnaire at their convenience following class. The researcher provided an overview of each section of the questionnaire (https://www.canva.com/design/DAEpt0euNQU/yZXGm4kQgbQaCVvyPubsNg/view?utm) and with the guidance provided, the learners filled it out accordingly.

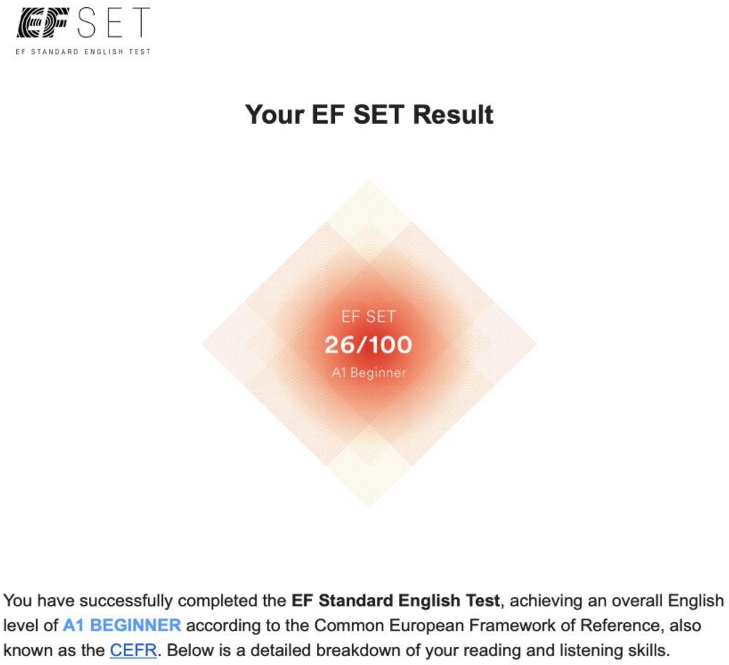
[46] The production task demonstrated a screenshot from a production training method utilised by (Ding et al., 2019). Participants were informed that while the fundamental structure remained consistent - recording their voices, listening to a model speaker, and then hearing their own voices again - the content varied. Unlike Ding et al. (2019), which used sentences, the current study used a word list.

Learners were given an information sheet in which they were briefly informed about the necessary details regarding the training (see Appendix A). They were then required to sign a consent form before participating in the study; they kept a copy for their records.

In week 3, all learners performed A 50 MIN European Framework Standardised English Test (EF SET) online (https://www.efset.org/ef-set-50/) since it was necessary to assess their language competence levels before the start of the training programme. The test is free and fully conforms to the Common European Framework of Reference (CEFR). It aligns with the CEFR's six levels of foreign language proficiency: A1 Beginner, A2 Elementary, B1 Intermediate, B2 Upper Intermediate, C1 Advanced, and C2 Proficient. 112 learners, categorised as A1 to represent their beginner level in English speaking, made up the primary sample for the study. To guarantee that all participants fully grasped the test components, the researcher organised a Zoom meeting. During this meeting, a PowerPoint presentation was shown, featuring screenshots of the test's two primary sections (listening and reading) and instructions on how to share the test results (https://www.canva.com/design/DAF5A3lRI2I/9Iv5WuiIvU2VKqF6I2T6iw/edit). It was made clear that after finishing the test, the score of each participant would be shown (see the sample score provided). They were instructed to capture a screenshot of their score and send it via email to the researcher, along with the username they intend to utilise throughout the training and testing sessions[48].

[48] Participants were also required to provide their usernames and EF SET test scores into a Google spreadsheet, providing a simple and efficient way to assess their proficiency levels.

\



Your EF SET Result

EF SET
26/100
A1 Beginner

You have successfully completed the **EF Standard English Test**, achieving an overall English level of **A1 BEGINNER** according to the Common European Framework of Reference, also known as the CEFR. Below is a detailed breakdown of your reading and listening skills.

Figure 3.4 A sample score of EF SET.

Prior to beginning the training sessions, learners were given a pre-test in week 4 to evaluate their perception of vowels. Learners conducted two training sessions[49] per week (from week 5 through week 8) as part of the first training phase[50]. After eight training sessions, learners took a mid-test in week 9 to evaluate their progress. Then, from weeks 10 to 13, they began the programme's second phase. Learners took a post-test in week 14 to assess their performance after completing all training sessions. The first generalisation test was conducted during week 15, whereas the second generalisation test took place during week 16, to evaluate the generalisation to novel stimuli.

[49] Each training session includes a variety of perceptual tasks including production, auditory discrimination, category discrimination, identification and the LingLab vowel-matching game (See section 2.4.5 for more details).

[50] Each training task was preceded by trials which were only shown in the first three sessions to aid participants in adjusting to the tasks.

Although the training is lengthy, spanning 16 sessions over three months, it is less intensive compared to previous studies. For instance, Iverson et al. (2005) held 10 sessions across 2-3 weeks, and Iverson et al. (2012) conducted 8 sessions over 1-2 weeks. This training's bi-weekly sessions were chosen to align with the preferences of most learners, effectively accommodating their academic and personal schedules by planning sessions for weekends (Fridays and Saturdays). Adapting the frequency of sessions to their availability helps ensure that participants can fully engage with and successfully complete the training. The training sessions were conducted with no more than one daily session to ensure consistent exposure (e.g., Iverson & Evans, 2009). For instance, session 1 was held on Friday and session 2 on Saturday of week 1. This approach was applied to all 16 sessions to maximise training effectiveness by preventing cognitive overload and allowing sufficient time for information processing between sessions.

To increase learner engagement and commitment during the training course, the researcher arranged brief weekly Zoom meetings on weekends, held just before the sessions started. Each meeting lasted between 10 to 15 minutes. The purpose of the meetings was not only to address learners' inquiries or concerns regarding the tasks and tests but also to consistently provide support and motivation for completing the training. The training was delivered online, so these meetings were crucial for the learning experience. They ensured ongoing interaction and provided necessary clarifications, thus contributing to a more effective and interactive training process. This section provides a thorough overview of the population description and sampling, the mode of training delivery, and the training plan's framework. The following section delves into the rationale for the training length, the incorporation of variability into the training, an overview of the training and testing tasks, the design of the introductory phase as well as the design of tasks used in both the training and testing sessions.

3.6 Design

3.6.1 Length and number of training sessions

To address the research questions, a pretest-posttest experiment was designed. Sixteen 45-minute training sessions were conducted over 8 weeks (two sessions per week). The study used the maximum number of sessions (15) shown to be advantageous for learners (Iverson & Evans, 2009) plus one additional session to balance the identification and discrimination tasks, for a total of 16 sessions. According to Iverson and Evans (2009), while Spanish learners' vowel perception did not improve after five sessions of the HV training program, German learners did. However, after completing ten additional sessions of the technique, Spanish learners achieved a similar level to German learners in a subsequent experiment. The researchers suggested that German learners' relatively crowded vowel space may have facilitated rather than hindered vowel learning. Spanish learners may take longer to learn English vowels due to their sparse L1 vowel system. They retained, however, the same fundamental capacity for learning once an additional ten training sessions were completed, bringing the total to fifteen. Similarly, Arabic learners have a sparse vowel system, requiring a longer training period to ensure learning.

3.6.2 Variability

The current study provides variability in the following ways:

1- **Variability of talkers/ accents**

The study incorporated a range of English varieties for learners, predicting a possible advantage of exposure to accent variation. That is, Arabic language learners who would expose to a variety of English accents would do better at identifying, discriminating, and producing L2 vowels than those exposed to a single accent. If typical HV training approach (i.e., acquiring knowledge from past encounters with multiple speakers and phonetic contexts) aids learning (enhance the cognitive processing of unfamiliar speakers and linguistic inputs) (e.g., Logan et al.,1991; Bradlow & Bent, 2003), then HV accent training should be beneficial for learners, and would support the generalisation to novel accents. The design of group A training paradigm was compatible with the perceptual design of the majority of HV methods described in the literature (e.g., Lively et al., 1993; Iverson & Evans, 2009; Grenon et al., 2019), all of which used a single homogeneous accent, often SSBE. On the other hand, the design of group B and C training programs applied in this study were evaluated to determine if exposure to multiple varieties is helpful for FL learners. Regardless of the accents employed, the three training programs (A, B, C) used multiple speakers and genders (2 female, 1 male) rather than a single speaker design. The difference is that group B and C training programs used one speaker of three accents for a total of three speakers, whereas group A used multiple speakers of the same accent.

2- Contextual variability in phonetics/ linguistics

The same stimuli were utilised in all three training conditions (HV-S, HV-M1, HV-M2). Vowels were heard in various consonantal contexts (e.g., h_d, b_t, b_n, f_m, m_t, f_l, m_t, b_k, l_k, b_d, k_d, p_t, r_st, k_l, l_d, sk_t, w_l, g_t, b_k, r_t, w_d, t_n, b_l, t_t, h_t, st_, k_k). The primary criterion for selecting stimuli for the training and testing phases was to include as many real words as possible, corresponding to words commonly heard in real interactions. However, a limited number of nonsense words were used for the vowels /ʊ/, /uː/, /ɛə/, /ɜː/, and /ɔɪ/, with the rationale for this explained in section 3.6.3. The testing and training stimuli are thoroughly described in Appendix D and E.

3.6.3 Overview of the training and testing tasks

Each training session consisted of five tasks: A forced-choice identification task, an auditory discrimination task, a category discrimination task, a production task, and the LingLab vowel matching game (RSE team, 2021). The following section describes how each of these tasks was designed. According to the reviewed research, HV identification and discrimination training are useful methods for improving generalisation and learning retention (e.g., Carlet & Cebrian, 2019; Shinohara & Iverson, 2018; Wayland & Li, 2008; Flege, 1995). For example, in a study conducted by Shinohara and Iverson (2018), it was found that both HV identification and discrimination training yielded equally effective results, leading to comparable levels of improvement. The training approaches improved learners' abilities in identification, auditory discrimination, category discrimination, and the production of the /l/-/r/

contrast. According to these findings, training in HV identification and discrimination is expected to help Arabic-speaking learners improve their perception of English vowels. In addition, incorporating different perceptual tasks was intended to further improve learners' engagement, particularly given the longer duration of the training.

It was decided to begin with the production task (PROD) followed by two discrimination tasks (AD, CD), an identification task (ID), and end with a card memory game. The purpose of beginning each session with a modelled production activity was to familiarise learners with the target vowels. Reproducing the words after hearing them from a model speaker prepared learners to discriminate and identify the vowels later. Integrating production and perception training has been shown to improve FL learners' productive abilities more effectively than either perception or production training alone (e.g., Alshangiti, 2015; Wong, 2014). These studies also indicate that hybrid training does not hinder perceptual improvement, as it leads to comparable gains in perceptual abilities as perception-focused training. Consequently, it is anticipated that the combined approach (i.e., HV perception and production training) would help Arabic learners enhance their perceptual skills. While the current study does not examine the combined approach's impact on perception and production, it adopts this method strategically, drawing on existing evidence that suggests it is beneficial for learners and does not hinder perceptual improvement. This decision is further reinforced by the SLM-r model, which posits a bidirectional relationship between production and perception, where advancements in one domain enhance the other, promoting simultaneous development. The SLM-r highlights the dynamic and interactive nature of this process, emphasising the mutual influence of these skills throughout the language language learning.

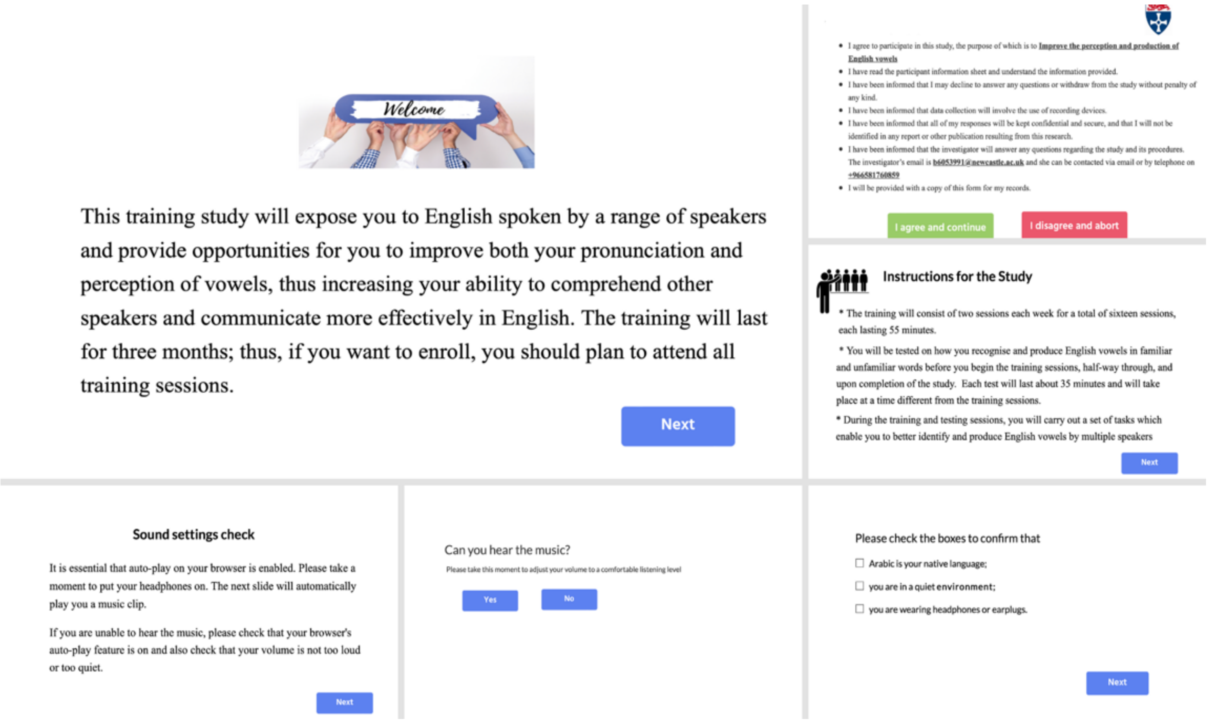
Five tests were administered throughout the training programme: a pre-test, a mid-test, a post-test, and two generalisation tests. Except for The LingLab vowel matching game, the tasks utilised in these tests were identical to those used in the training session. None of the tests had feedback, which was only supplied during training sessions. The table below outlines the tasks performed during the training and testing phases, while the following section describes how to design each task.

	Tasks employed throughout the experimental groups (A, B, C)
Pre-test: **Week 4**	• Identification task— No feedback • Auditory discrimination— No feedback • Category discrimination— No feedback • Production task — reading word lists and a passage— No feedback
Training part 1 **Weeks 5-8**	• Identification task— with immediate feedback • Auditory discrimination— with immediate feedback • Category discrimination— with immediate feedback • Production task— Feedback → produce- listen- record • The LingLab vowel-matching game
Mid-test: **Week 9**	Same procedures as the pre-test
Training part 2 **Weeks 10-13**	Same procedures as training part 1
Post-test **Week 14**	Same procedures as the pre-test
Generalisation test I **Week15**	Same procedures as the pre-test
Generalisation test II **Week 16**	Same procedures as the pre-test

Table 3.5 Testing and training tasks used in the current study.

3.6.4 Designing the Welcome Start

Before administering the pre-test, an opening greeting was developed (see Figure 3.5 below). The first slide explained the training's main objectives. If participants wish to continue the training, they should click "I agree and continue" on the consent form displayed on the next screen. If they did not choose to proceed, they could click the "*I disagree and abort*" button. Following that, the study instructions were presented on a slide. The sound settings check was subsequently introduced, informing the participant about the importance of enabling AutoPay and notifying them that the following slide would automatically begin music playback to verify the activation of the browser's auto-play functionality while ensuring that the volume was appropriately adjusted, neither too loud nor too soft. In the subsequent slide, each participant is required to confirm their native language as Arabic, ensure they are in a noise-free environment, and utilise headphones or earplugs.



This training study will expose you to English spoken by a range of speakers and provide opportunities for you to improve both your pronunciation and perception of vowels, thus increasing your ability to comprehend other speakers and communicate more effectively in English. The training will last for three months; thus, if you want to enroll, you should plan to attend all training sessions.

Next

- I agree to participate in this study, the purpose of which is to **Improve the perception and production of English vowels**
- I have read the participant information sheet and understand the information provided.
- I have been informed that I may decline to answer any questions or withdraw from the study without penalty of any kind.
- I have been informed that data collection will involve the use of recording devices.
- I have been informed that all of my responses will be kept confidential and secure, and that I will not be identified in any report or other publication resulting from this research.
- I have been informed that the investigator will answer any questions regarding the study and its procedures. The investigator's email is b6053991@newcastle.ac.uk and she can be contacted via email or by telephone on +966581760859
- I will be provided with a copy of this form for my records.

I agree and continue   I disagree and abort

Instructions for the Study

* The training will consist of two sessions each week for a total of sixteen sessions, each lasting 55 minutes.

* You will be tested on how you recognise and produce English vowels in familiar and unfamiliar words before you begin the training sessions, half-way through, and upon completion of the study. Each test will last about 35 minutes and will take place at a time different from the training sessions.

* During the training and testing sessions, you will carry out a set of tasks which enable you to better identify and produce English vowels by multiple speakers

Next

Sound settings check

It is essential that auto-play on your browser is enabled. Please take a moment to put your headphones on. The next slide will automatically play you a music clip.

If you are unable to hear the music, please check that your browser's auto-play feature is on and also check that your volume is not too loud or too quiet.

Next

Can you hear the music?

Please take this moment to adjust your volume to a comfortable listening level

Yes   No

Please check the boxes to confirm that

☐ Arabic is your native language;

☐ you are in a quiet environment;

☐ you are wearing headphones or earplugs.

Next

Figure 3.5 Designing the Welcome Start.

3.6.5 Designing the training tasks

The present study comprised 16 training sessions, each involving 108 trials across five tasks (ID, AD, CD, PROD, The LingLab vowel matching game), as detailed in Figure 3.6 below. Each session targeted specific subsets of stimuli across these tasks, except the vowel game. For details on how vowels were distributed throughout the training, refer to section 3.3.1. In the PROD task, trials were fixed, ensuring participants encountered different model speakers each time, adjusted based on their group assignment—either different speakers with the same accent or different speakers with various accents. Meanwhile, the trials for all other tasks were randomised and counterbalanced across the sessions.

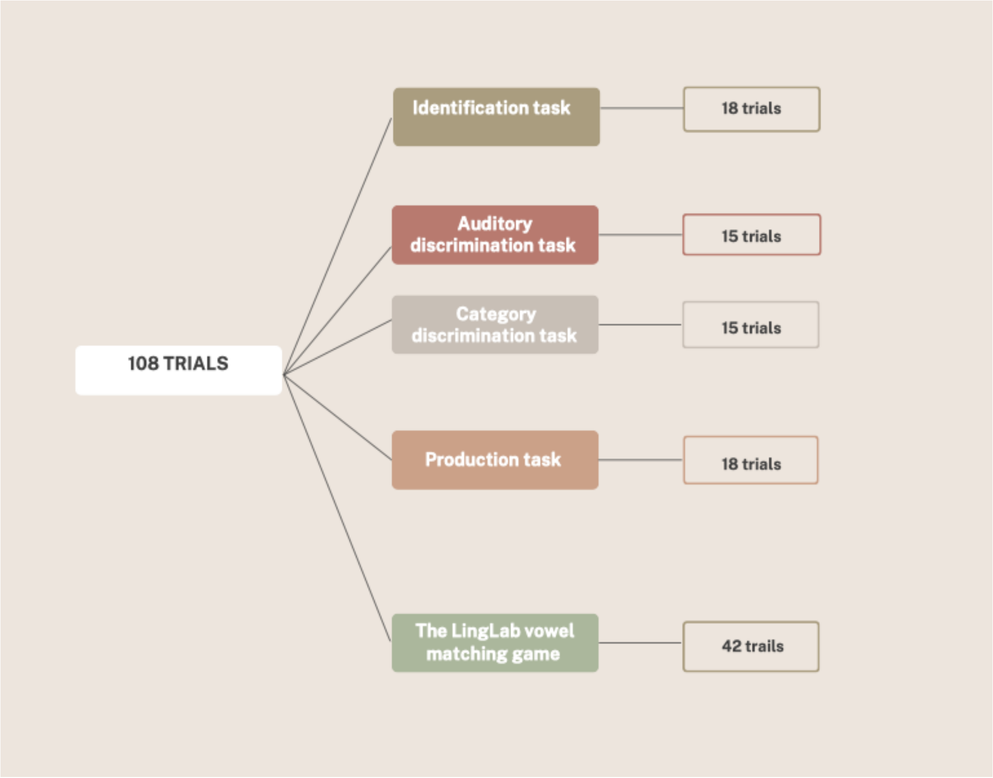
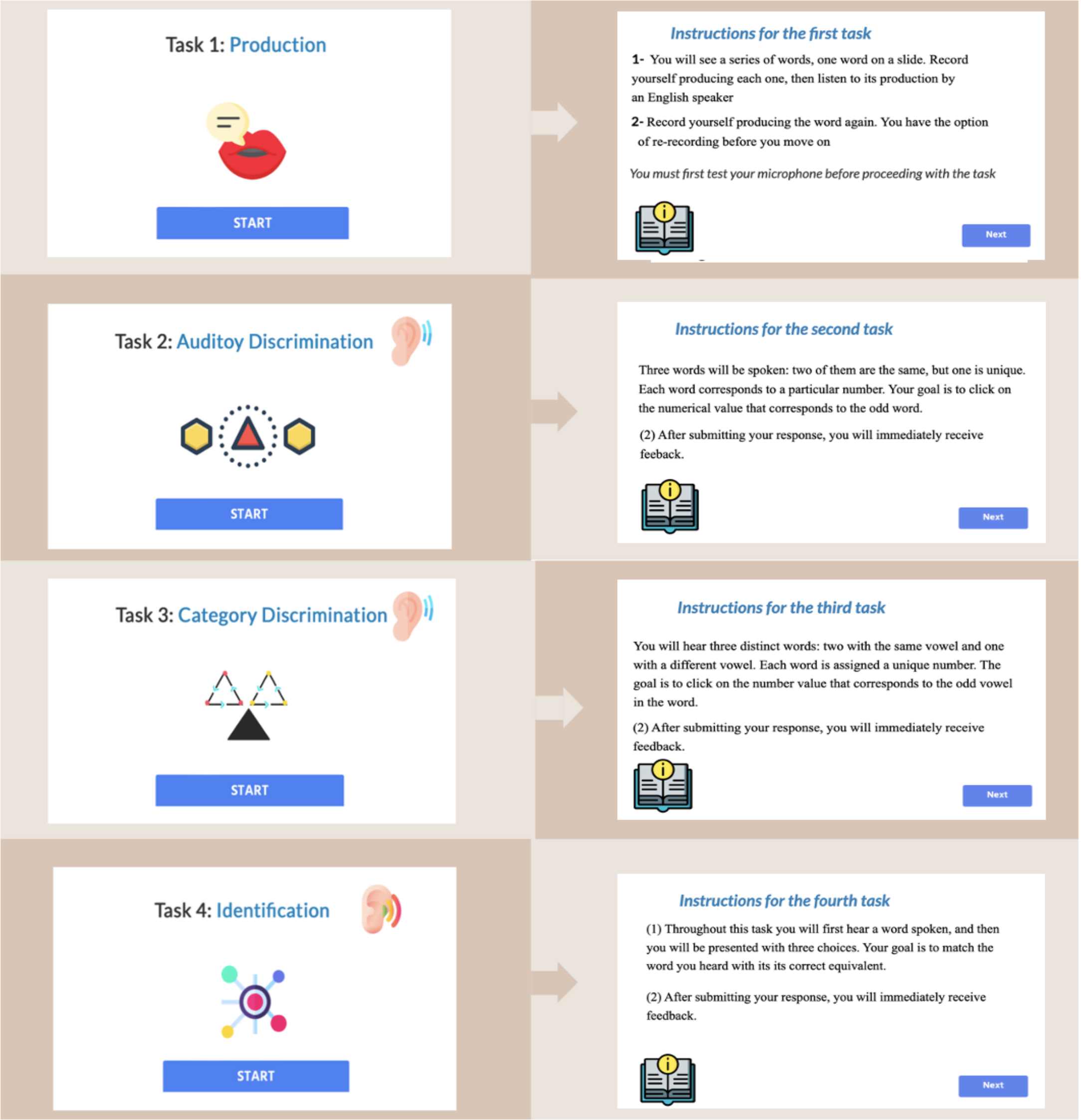


Figure 3.6 The number of trials used for each training tasks [production, auditory discrimination, category discrimination, identification, the LingLab vowel matching game.

The number of trials in this study (108) is modest compared to previous research, such as Iverson and Evans (2009) and Iverson et al. (2012), which included 225 trials in each of five HV identification sessions covering all target vowels. Despite fewer trials, the design is advantageous for beginners as it avoids cognitive overload that could result from presenting all stimuli simultaneously. This approach should promote a gradual learning process by breaking the material into manageable segments. Moreover, with the training extending over 16 sessions, spreading the stimuli across these sessions helps to mitigate potential boredom. Although this study has fewer trials than previous research (e.g. Iverson & Evans, 2009; Alshangiti, 2015), it ensures adequate exposure by presenting each vowel cluster or pair three times, each in different words and by different speakers, regardless of whether they share the same accent or have differing accents. It would be worthwhile to examine the effects of increasing the number of training trials on beginner learners in future studies while adhering to the same study design.

Before each task, three practice trials were designed for the initial three training sessions to help participants become accustomed to the tasks (refer to Appendix E). These trials were not part of the training or testing materials. As a reminder, textual instructions were provided in English prior to each task as shown in the figure below. Considering the beginner status of the participants, detailed explanations in Arabic were provided during the introductory sessions, which took place both in person and online at the programme's start. Furthermore, Arabic recorded materials explaining the tasks were made accessible on the Telegram channel as a reference resource. This strategy was crucial in ensuring that all participants fully comprehended the tasks, thus laying a solid foundation for their learning journey.



Task 1: Production

Instructions for the first task

1- You will see a series of words, one word on a slide. Record yourself producing each one, then listen to its production by an English speaker

2- Record yourself producing the word again. You have the option of re-recording before you move on

You must first test your microphone before proceeding with the task

Task 2: Auditoy Discrimination

Instructions for the second task

Three words will be spoken: two of them are the same, but one is unique. Each word corresponds to a particular number. Your goal is to click on the numerical value that corresponds to the odd word.

(2) After submitting your response, you will immediately receive feeback.

Task 3: Category Discrimination

Instructions for the third task

You will hear three distinct words: two with the same vowel and one with a different vowel. Each word is assigned a unique number. The goal is to click on the number value that corresponds to the odd vowel in the word.

(2) After submitting your response, you will immediately receive feedback.

Task 4: Identification

Instructions for the fourth task

(1) Throughout this task you will first hear a word spoken, and then you will be presented with three choices. Your goal is to match the word you heard with its its correct equivalent.

(2) After submitting your response, you will immediately receive feedback.

Figure 3.7 Task instructions provided to participants for production, auditory discrimination, category discrimination, and identification.

The tasks (PROD, AD, CD, and ID) were developed by the researcher using Labvanced. While designing these tasks did not require coding skills, technical issues and inquiries often arose, all of which were resolved by the Labvanced support team. Participants performed these tasks using a web browser, with links provided each week. Responses were collected through the browser, ensuring participants could not alter their answers and thus maintaining the integrity and accuracy of the data. The LingLab vowel matching game was designed with the assistance of the Research

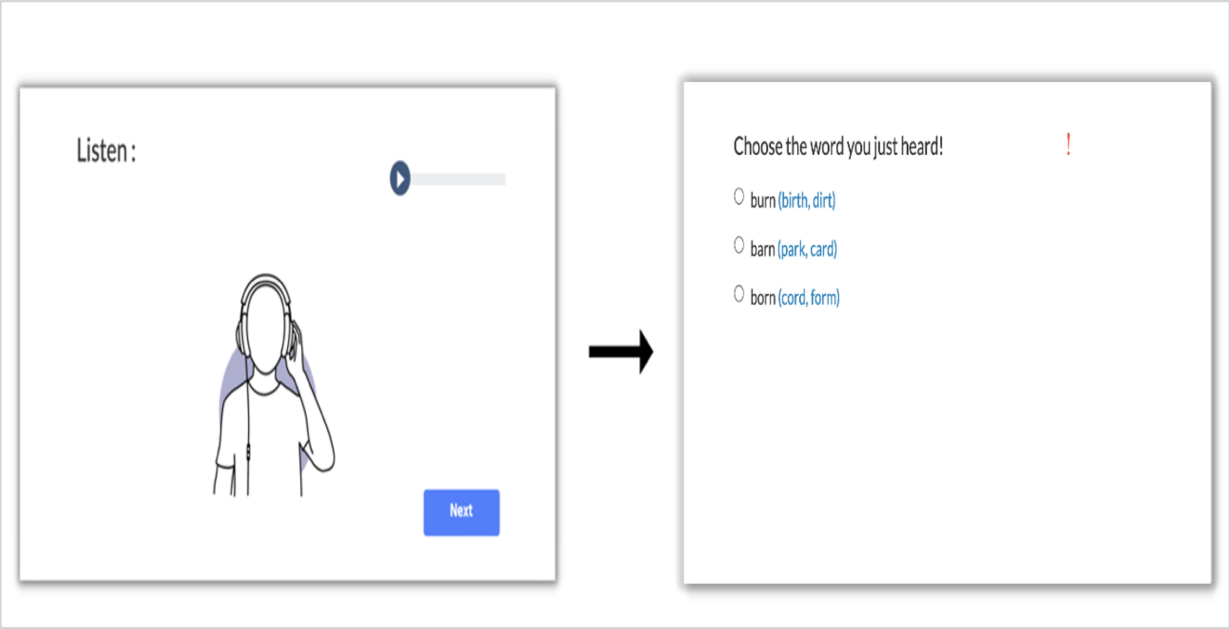
Software Engineering (RSE) Unit at Newcastle University and under the guidance of professor Ghada Khattab (RSE, 2021). Similarly, the game was performed using a web browser. Due to the time constraints for support available from the RSE team, they were unable to include the functionality to monitor progress, which required users to establish an account and log in. Instead, participants recorded their progress by filling out a supplied Google spreadsheet after each training session with their username, the length of time, and the number of moves spent on each game attempt (see section 3.6.5.5 for further details). The following subsections illustrate the design of each training task.

3.6.5.1 Identification

In this task, a single slide played a word-specific audio recording automatically. Subsequently, three minimal pairs were displayed on the slide, and participants were asked to identify the spoken word by selecting its written equivalent from these pairs. To guarantee that the learners were familiar with the stimuli throughout the training sessions, each response was accompanied by two common words (Iverson & Evans, 2009; Iverson et al., 2012). For example, after playing the stimulus 'born', the following slide presented 3 choices:

burn (birth, dirt)
barn (park, card)
born (cord, form)

The task for the participants was to identify which of these words corresponded to the spoken word 'born', Consider Figure 3.8 for the design of the identification task.



Choose the word you just heard! !

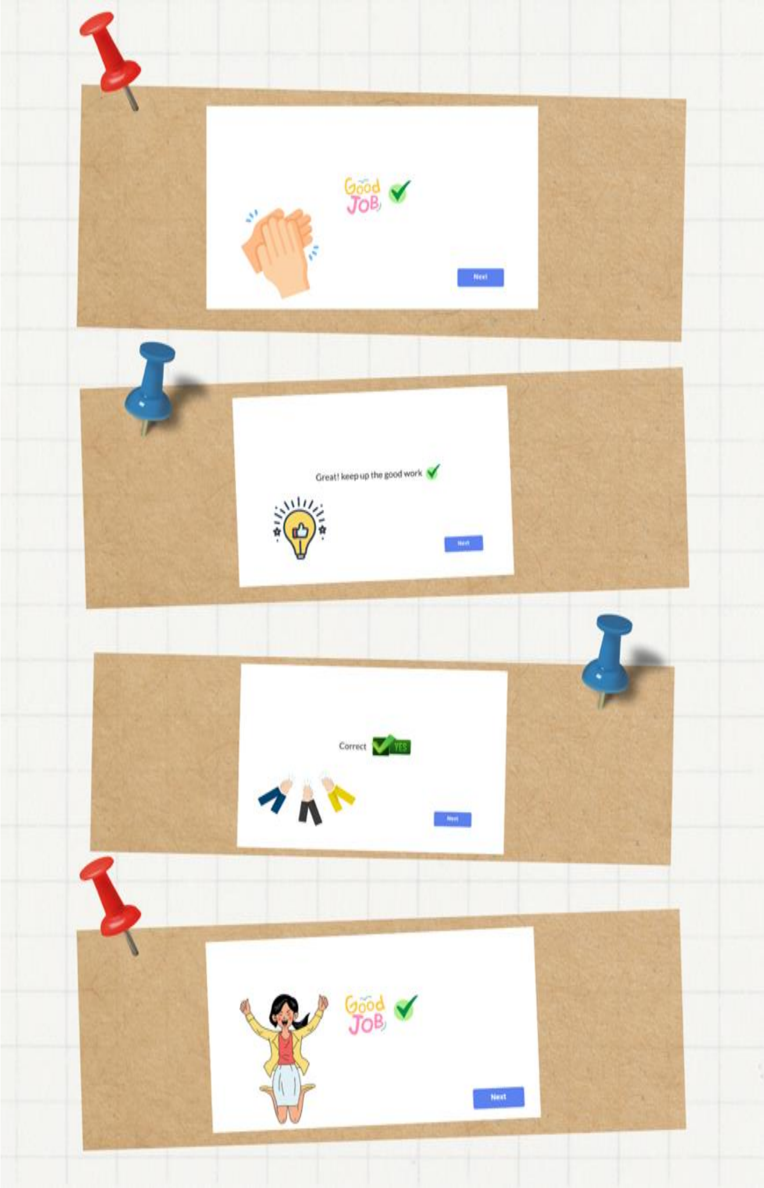
○ burn (birth, dirt)

○ barn (park, card)

○ born (cord, form)

8Figure 3. The design of the identification task

If the participants correctly identified the stimulus they heard (e.g., born) they received affirmative feedback on the succeeding slide. To boost interest in the activity, the representation of positive feedback was altered eight times during sixteen sessions. Figure 3.9 below represents a sample of the feedback provided to participants who selected the correct response.

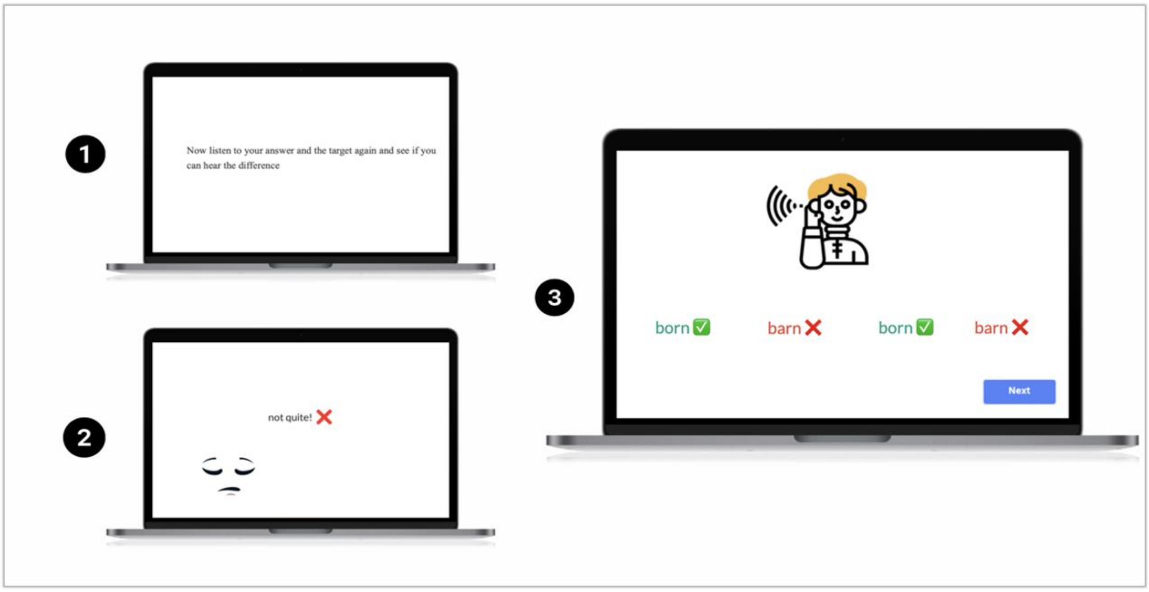


Great! keep up the good work ✓

Correct ✓ YES

Figure 3.9 Feedback provided for correct responses in the Identification task.

If participants selected the incorrect option, they heard a four-stimulus alternating series of the correct word, the incorrect response, the correct word, and the incorrect response, along with orthographic representations where correct words were in green and the incorrect response in red. During the introductory sessions, all participants were instructed that the correct response would always play first, followed by their selected incorrect response. For example, if the stimulus 'born' played while the participant selected the incorrect answer ‘barn’ from the options ‘burn, barn, born,’ four audio recordings were played: ‘born, barn, born, barn,’ beginning with the correct response and then the incorrect selected response (see Figure 3.10 below). Ideally, participants should first hear the correct word played, followed by a four-stimulus alternating series of the correct and incorrect words (e.g., Iverson et al., 2012). However, as participants were fully informed about the order of correct and incorrect responses through the introductory sessions and practice trials, the feedback remained effective.



Now listen to your answer and the target again and see if you can hear the difference

not quite! ✗

born ✓ barn ✗ born ✓ barn ✗

Figure 3.10 Feedback for incorrect responses (Identification task).

3.6.5.2 Auditory Discrimination

To establish auditory discrimination, each trial consisted of the presentation of three audio files, two of which included similar words and the third of which contained a unique word. The odd word was randomly placed in one of three slots. All audio files were invisible and automatically played once. The inter-stimulus delay between words was 200 ms, which was informed by prior research (Gerrits 2001). Three images were displayed in the centre of the screen for each trial. Each of the three images represented a spoken word and one of the digits 1, 2, or 3. Participants were required to hit the relevant number key below the image to identify the odd word (see Figure 3.11). The same speaker always uttered the three words in each trial; however, the speaker varied between trials. Note that the images were matched to the gender of the speaker. The order of trials was arbitrary every time a subject did the task.

Participants encountered different speakers based on the training approaches assigned to them.

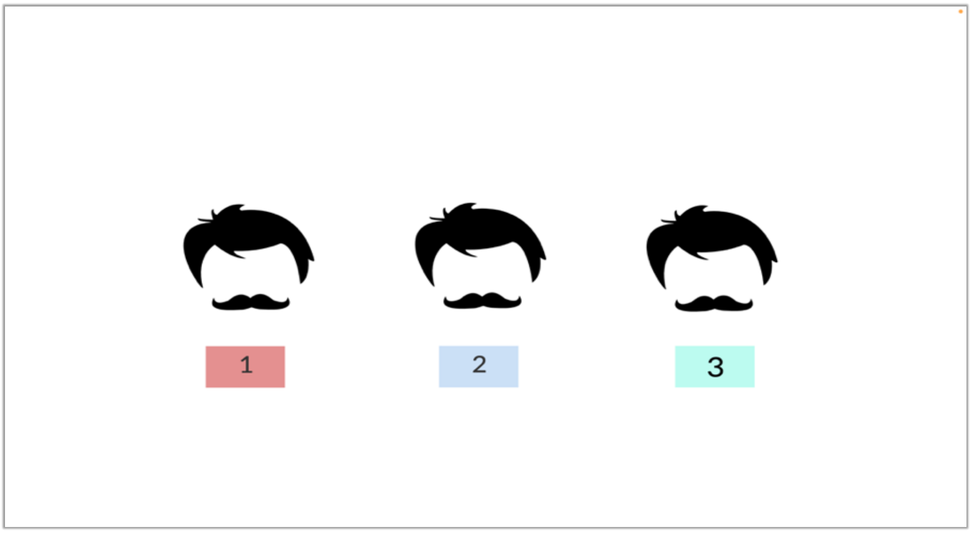
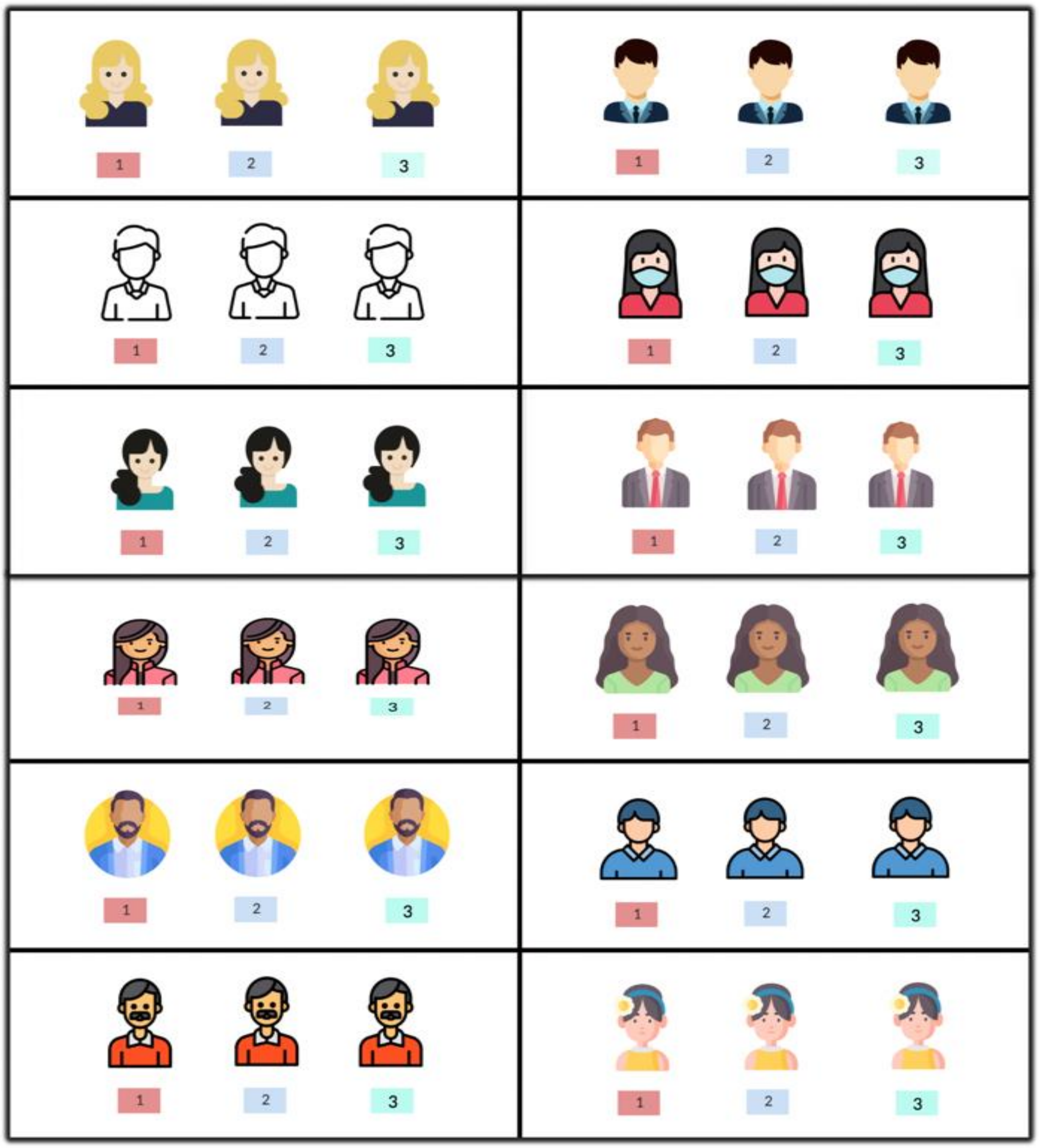


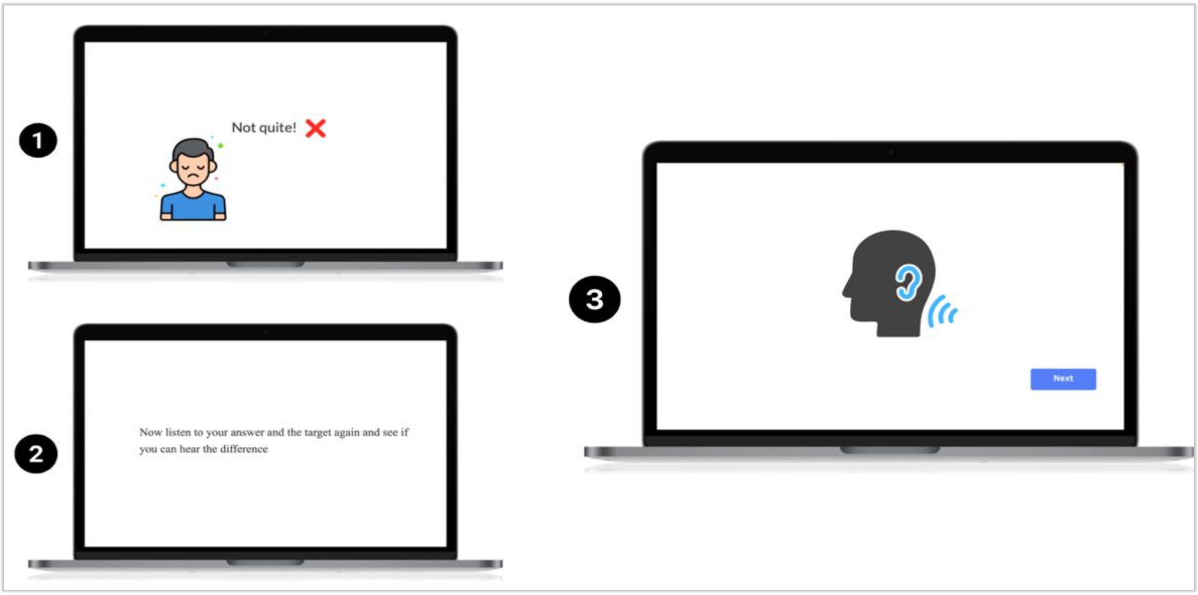
Figure 3.11 The design of the discrimination task.

Similar to the production task, the pictures used in the trials were updated every two sessions to make the training more engaging. Figure 3.12 illustrates a sample picture presentation.



Different image designs used in the discrimination task sessions to prevent boredom. **12 Figure 3.**

When participants chose the correct answer, they received positive feedback confirming their decision. However, if participants selected an incorrect response, they were directed to a slide that read, "*Now listen to your answer and the target again to determine if you can hear the difference*," followed by a four-stimulus alternating series of the correct word, the incorrect response, the correct word, and the incorrect response (Iverson & Evans, 2009) (see Figure 3.13). For example, if the participant selected 'met' from the options 'met, mitt, met,' four audio recordings were played: 'mitt, met, mitt, met,' always beginning with the correct response and then the selected response. Participants fully understood the order of correct and incorrect answers through the introductory sessions before the start of the training and the practice trials they completed during the initial three training sessions.



Not quite! ✗

Now listen to your answer and the target again and see if you can hear the difference

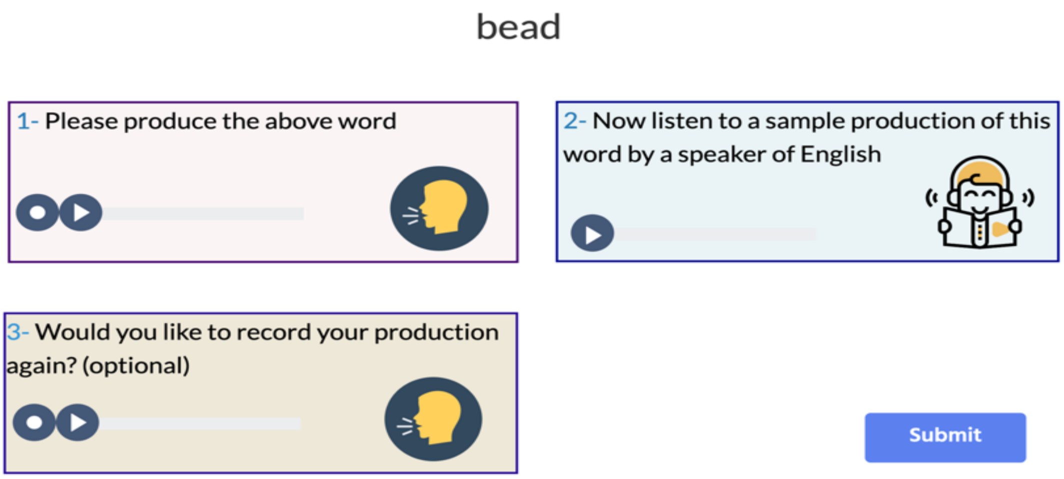
Figure 3.13 Feedback provided for incorrect responses in the auditory discrimination task.

3.6.5.3 Category Discrimination

Each trial of the category discrimination task consisted of presenting three audio files comprising three distinct words, two of which had the same vowel phoneme and one of which had a vowel phoneme different from the other two. Participants were required to select the word with the odd vowel phoneme. For example, if they heard the words ‘hit, bid, met’, they would be expected to choose ‘met’ because of its different vowel phoneme compared to the other two. This task's design, presentation, and feedback were identical to the preceding auditory discrimination task (see section 3.4.5.2).

3.6.5.4 Production

In this task, three boxes were displayed below each presented word (see Figure 3.14 below). The *first box* required the participant to record the word by clicking the recording button. When they pressed the play button, their recording began to play. If they wished to record the word a second time, they could hit the record button again. In the *second box*, a model speaker's recorded pronunciation of the word was added. The provided audio files varied depending on the group's classification. Table 3.6 depicts the classification of audio files based on the fixed trial order of the HV-S, HV-M1, and HV-M2 speaker groups. To prevent boredom, the designs of the three boxes were changed every two training sessions, both in terms of their colours and the images contained within them (observe some of the designs in Figure 3.15).



bead

1- Please produce the above word

2- Now listen to a sample production of this word by a speaker of English

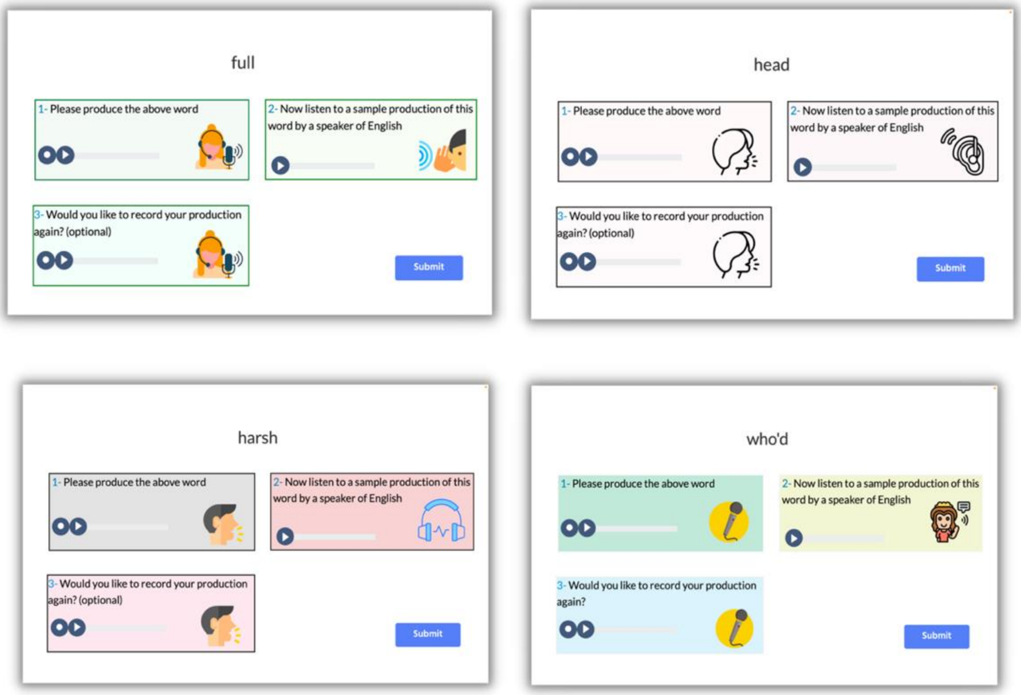
3- Would you like to record your production again? (optional)

Figure 3.14 The design of the production task.

To allow self-evaluation, all participants were told to produce each word before hearing it from a model speaker. This feedback (self-monitoring and self-correction) aimed to activate the learners' feedback sensory systems, allowing them to make fast motor adjustments in reaction to detecting significant differences between the speaker model and their output (Baker & Trofimovich, 2006). Participants could replay both their own recordings and the model talker’s examples to discern the differences (cf. Alshangiti, 2015). For a comprehensive list of the stimuli used throughout the 16 training sessions in the production task, refer to Appendix F.

Index	Trial Id (fixed)	Audio files	HV-S	HV-M1	HV-M2
1	1	Word 1	UK speaker 1	UK speaker 1	UK speaker 1
2	2	Word 2	UK speaker 2	US speaker 2	US speaker 2
3	3	Word 3	UK speaker 3	AUS speaker 3	SA speaker 3
4	4	Word 4	UK speaker 1	UK speaker 1	UK speaker 1
5	5	Word 5	UK speaker 2	US speaker 2	US speaker 2
6	6	Word 6	UK speaker 3	AUS speaker 3	SA speaker 3
7	7	Word 7	UK speaker 1	UK speaker 1	UK speaker 1
8	8	Word 8	UK speaker 2	US speaker 2	US speaker 2
9	9	Word 9	UK speaker 3	AUS speaker 3	SA speaker 3
10	10	Word 10	UK speaker 1	UK speaker 1	UK speaker 1
11	11	Word 11	UK speaker 2	US speaker 2	US speaker 2
12	12	Word 12	UK speaker 3	AUS speaker 3	SA speaker 3
13	13	Word 13	UK speaker 1	UK speaker 1	UK speaker 1
14	14	Word 14	UK speaker 2	US speaker 2	US speaker 2
15	15	Word 15	UK speaker 3	AUS speaker 3	SA speaker 3
16	16	Word 16	UK speaker 1	UK speaker 1	UK speaker 1
17	17	Word 17	UK speaker 2	US speaker 2	US speaker 2
18	18	Word 18	UK speaker 3	AUS speaker 3	SA speaker 3

Table 3.6 Speakers heard throughout the production task.



full

1- Please produce the above word

2- Now listen to a sample production of this word by a speaker of English

3- Would you like to record your production again? (optional)

head

1- Please produce the above word

2- Now listen to a sample production of this word by a speaker of English

3- Would you like to record your production again? (optional)

harsh

1- Please produce the above word

2- Now listen to a sample production of this word by a speaker of English

3- Would you like to record your production again? (optional)

who'd

1- Please produce the above word

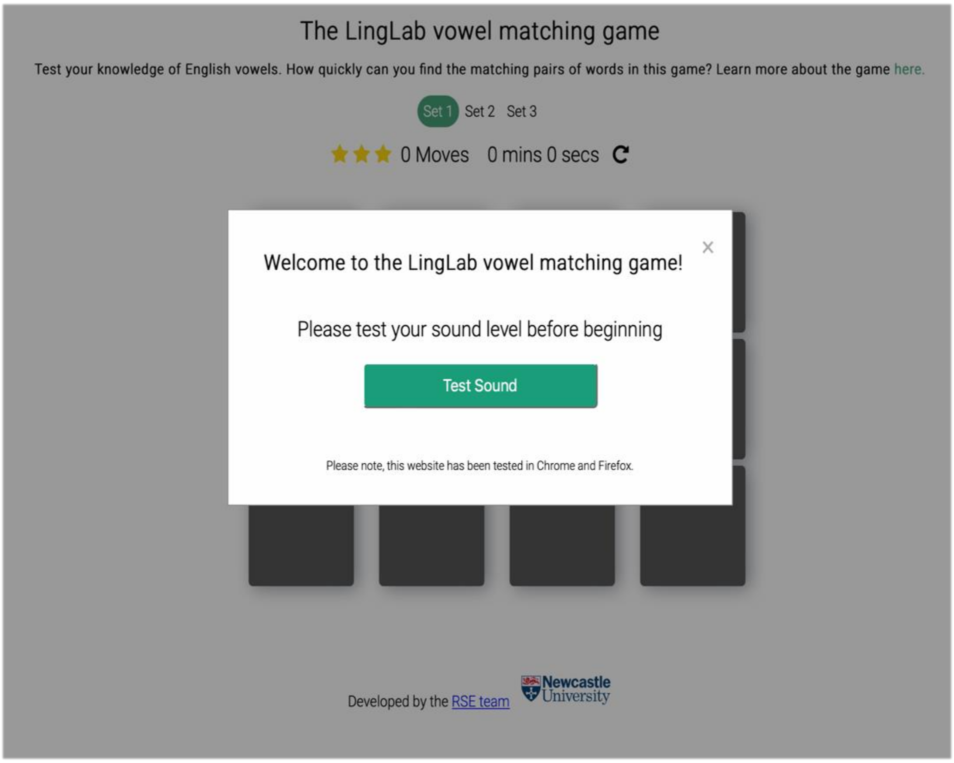
2- Now listen to a sample production of this word by a speaker of English

3- Would you like to record your production again?

Figure 3.15 Various formats used for the production task to prevent boredom.

4.6.5.5 The LingLab vowel matching game

The design of LingLab vowel matching game (RSE, 2021) was inspired by the UCL memory card game (Iverson et al., 2023), which incorporated all necessary training components with high variability (multiple talkers, multiple phonetic contexts). It was adapted for this thesis by including multiple English varieties. It was utilised as a part of a variety of training activities intended to enhance Arabic learners' perception of English vowels. When clicking the game's URL, a pop-up box demanding a sound test would appear (see Figure 3.16). After adjusting the volume to a comfortable level, Sets 1, 2, and 3 were displayed. **Set 1** included three speakers with British accents. **Set 2** consisted of three speakers with British, American, and Australian accents, while **Set 3** consisted of three with British, American, and Saudi accents. The decision to incorporate three sets was made to ensure consistency with the design of the previous tasks (created using Labvanced), particularly in relation to the accents that participants were exposed to. As a result, group A was assigned to play the game with Set 1, group B was assigned to play the game with Set 2, and group C played the game with Set 3.


The LingLab vowel matching game

Test your knowledge of English vowels. How quickly can you find the matching pairs of words in this game? Learn more about the game here.

Welcome to the LingLab vowel matching game!

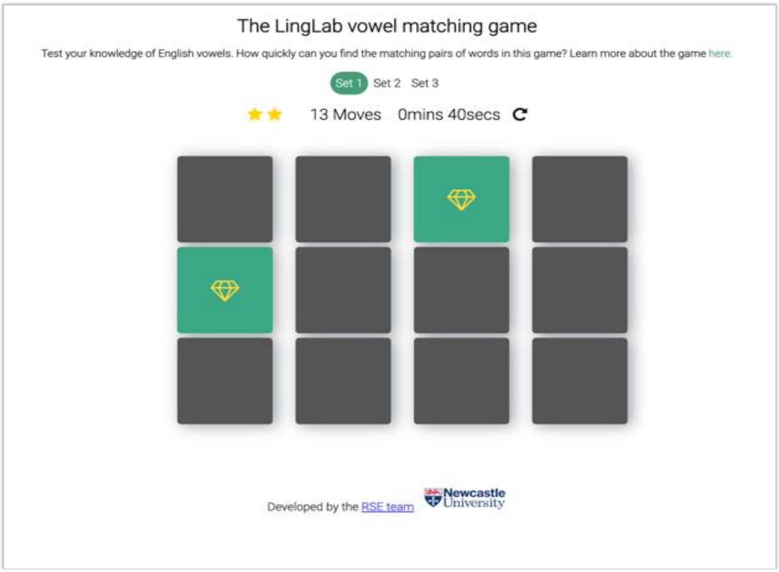
Please test your sound level before beginning

Please note, this website has been tested in Chrome and Firefox.

Developed by the RSE team

Adjusting the volume for the Linglab vowel matching game.**16 Figure 3.**

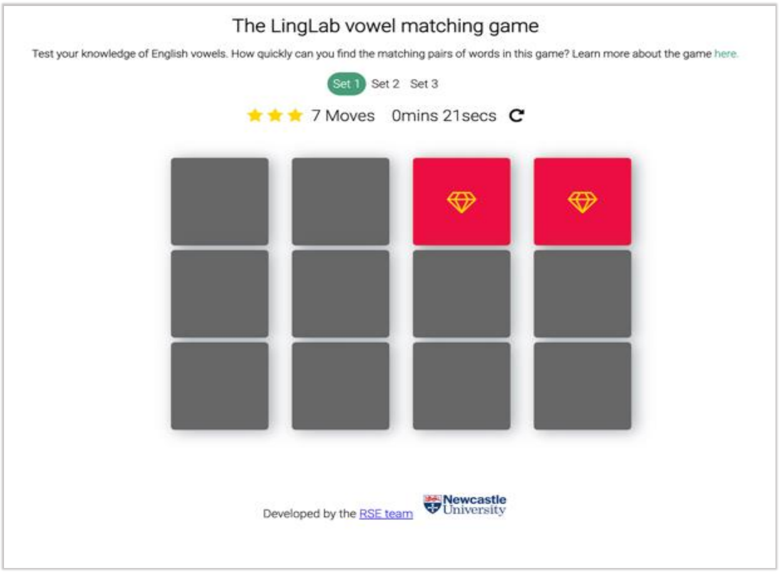
To play the game, participants were given explicit instructions to quickly and accurately find pairs of corresponding words, an approach designed to boost their focus and enhance learning. For instance, when the word 'mate' was played, participants needed to immediately search for and select the matching word from the displayed array of cards. Successful matches would cause the cards to turn green, signaling a correct pairing, while incorrect selections would turn the cards red, indicating a mismatch. Figures 3.17 and 3.18 depict illustrations of correct and incorrect card matches.



The LingLab vowel matching game

Test your knowledge of English vowels. How quickly can you find the matching pairs of words in this game? Learn more about the game here.

Figure 3. Correctly matched cards displayed in green.**17**

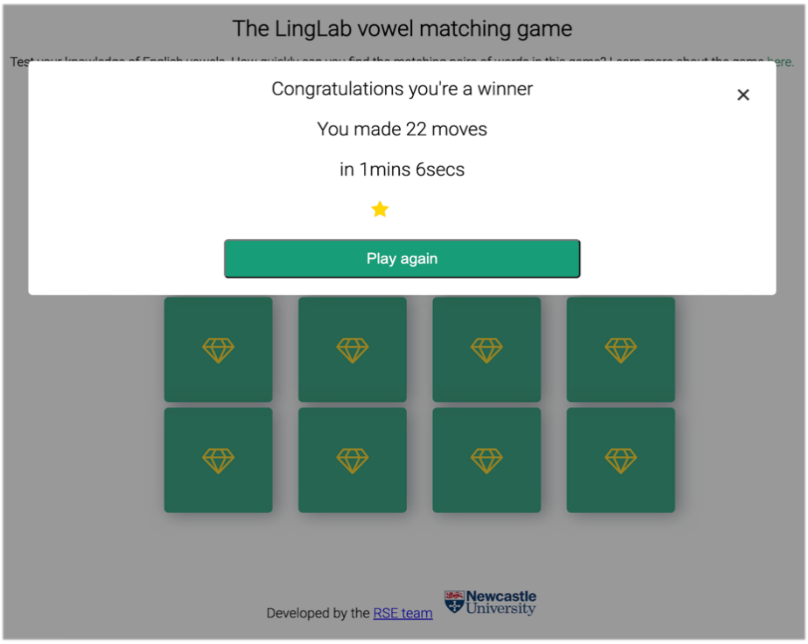


The LingLab vowel matching game

Test your knowledge of English vowels. How quickly can you find the matching pairs of words in this game? Learn more about the game here.

Figure Incorrectly matched cards displayed in red.**183.**

Once all cards were matched, the text "Congratulations! You're a winner! You made (X) moves in X min X secs. Play again" was displayed (A snapshot of the feedback is shown in Figure 3.19). Participants engaged in the game seven times, assuring an overall play duration of 10 to 15 minutes, essential for sustaining engagement and tracking play speed. If the time required to complete the game decreases with each playthrough, this indicates enhanced game performance. As previously mentioned, the RSE team could not include progress-monitoring functionality requiring user accounts due to time constraints. Instead, participants recorded their progress in a provided Google spreadsheet after each training session, noting their username, the length of time, and the number of moves per game attempt. One additional constraint of the game was that learners had access to Sets 1, 2, and 3. Nevertheless, detailed instructions were provided for each respective group: group A was exclusively instructed to utilise Set 1, group B to employ Set 2, and group C to use Set 3. To ensure strict adherence to this structure, the researcher consistently reinforced the learners' assigned sets during weekly online meetings.



The LingLab vowel matching game

Test your knowledge of English vowels. How quickly can you find the matching pairs of words in this game? Learn more about the game here.

Congratulations you're a winner

You made 22 moves

in 1mins 6secs

Feedback on the LingLab vowel matching game, showing the number of moves and time**19 Figure 3.** taken.

To assess the viability of the design, the researcher carried out an online pilot of training tasks with 11 Saudi Arabic speakers, including 6 English students and 5 computer science students. Learners found the design of the production task, the auditory discrimination task, and the LingLab vowel matching game to be effectively structured. However, they encountered difficulties with the category discrimination and identification tasks, implying that the stimuli should be automatically replayed twice rather than once. Despite this, the suggestion was not implemented since participants had been informed that while the tasks may appear challenging initially, they would become more manageable over time. In addition, it has been verified that the time intervals between stimuli in the AD and CD tasks are appropriate for the learners.

3.6.6 Testing procedures

The design of the perceptual tasks employed in the testing phase (pre-test, mid-test, post-test, gen1 test, and gen2 test) was identical to those used in the training phase (i.e., ID, AD, CD), except that participants did not receive feedback (See sections 3.4.5.1, 3.4.5.2, 3.4.5.3, 3.4.5.4).

Regarding the production test[51], there were two tasks:

- The first is a word-reading task in which participants were instructed to produce target words in a regular speaking tone (neither too loud nor too soft) in a quiet setting. Each slide displayed a word in the centre of the screen. Participants were required to click the recording button below the word. They could re-record the word by hitting the record button again. No model speaker was offered to obtain their spontaneous production (see Figure 3.20). Due to the variety of recording devices used by participants, they were instructed to record each word twice. This was to guarantee that at least one recording was successfully registered on the Labvanced website. Additionally, participants were advised to wait a few seconds after recording each word so that the audio files could be uploaded to the website before proceeding to the next slide.

[51] Production test data was collected for analysis following the completion of the PhD, indicating that this thesis does not cover the assessment of learners' production.

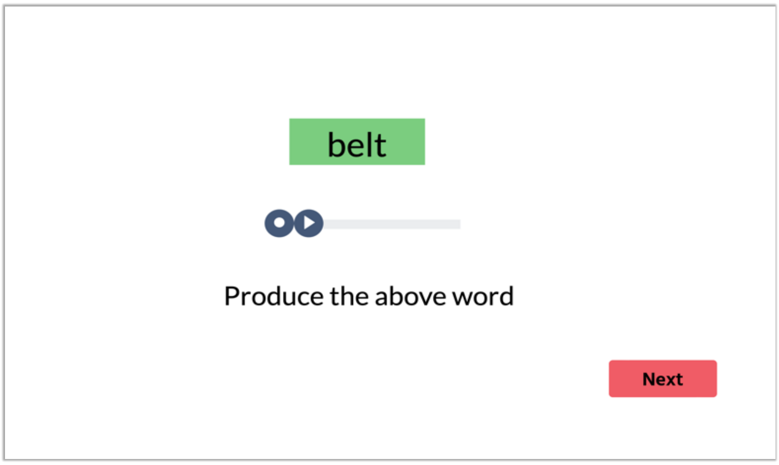
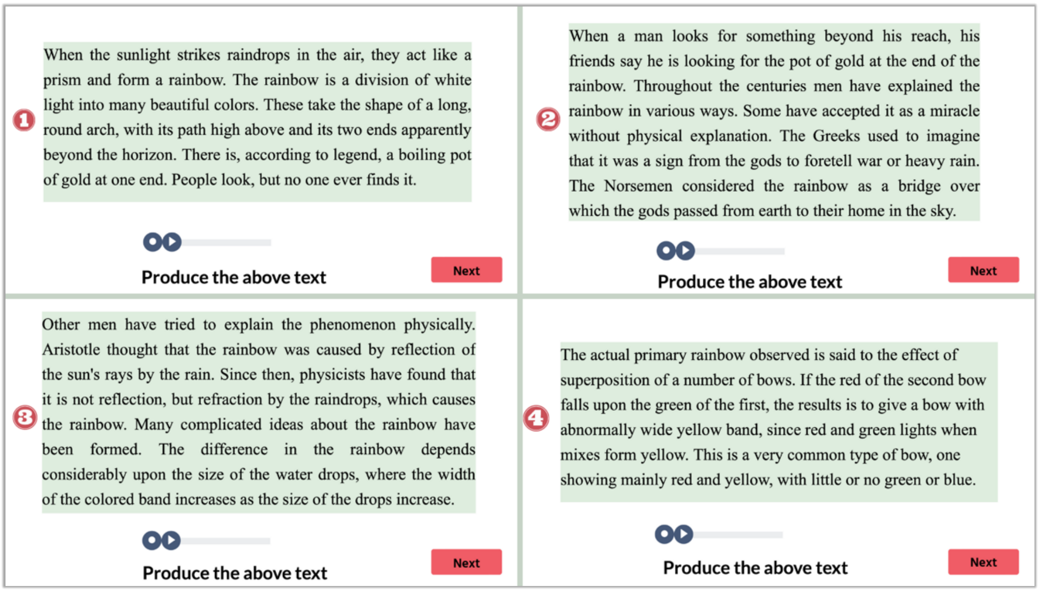


Figure 3.20 Designing the first production test

- The second task was a passage reading task in which participants read a short story including the study's target vowels on each test (see Appendix D). Similar to the word-reading activity, participants recorded the passage twice for each test to ensure delivery. The passage's layout was divided into four slides for ease of reading. Consider the layout of the 'rainbow' passage (Fairbanks, 1960) used in the pre-test, mid-test, and post-test, which is divided into four pieces to improve typographic readability (see Figure 3.21 below).



1 When the sunlight strikes raindrops in the air, they act like a prism and form a rainbow. The rainbow is a division of white light into many beautiful colors. These take the shape of a long, round arch, with its path high above and its two ends apparently beyond the horizon. There is, according to legend, a boiling pot of gold at one end. People look, but no one ever finds it.

2 When a man looks for something beyond his reach, his friends say he is looking for the pot of gold at the end of the rainbow. Throughout the centuries men have explained the rainbow in various ways. Some have accepted it as a miracle without physical explanation. The Greeks used to imagine that it was a sign from the gods to foretell war or heavy rain. The Norsemen considered the rainbow as a bridge over which the gods passed from earth to their home in the sky.

3 Other men have tried to explain the phenomenon physically. Aristotle thought that the rainbow was caused by reflection of the sun's rays by the rain. Since then, physicists have found that it is not reflection, but refraction by the raindrops, which causes the rainbow. Many complicated ideas about the rainbow have been formed. The difference in the rainbow depends considerably upon the size of the water drops, where the width of the colored band increases as the size of the drops increase.

4 The actual primary rainbow observed is said to the effect of superposition of a number of bows. If the red of the second bow falls upon the green of the first, the results is to give a bow with abnormally wide yellow band, since red and green lights when mixes form yellow. This is a very common type of bow, one showing mainly red and yellow, with little or no green or blue.

Figure 3.21 Designing the second production test

The production tests administered during the generalisation tests (gen1, gen2) adhered to a comparable structure to that employed in the pre-/mid-/post-tests, albeit with new word list stimuli and a different reading passage 'Comma gets a cure' (Honorof et al., 2000).

3.7 Ethical approval

Ethical approval was obtained from the School of Education, Communication & Language Sciences (ECLS) ethics committee and from Alwajh College, where the training and data collection took place. The data were kept anonymous, and participants were given pseudonyms so they could not be identified. All participants received adequate notification of their right to withdraw from the training programme. However, participants were strongly encouraged to complete the entire programme if they enrolled. As an incentive for learners to complete the training programme, bonus credits were given to at least three subjects, exclusively for those who completed the training. During the initial on-campus meetings, printed *information sheets* were handed out to all participating learners, intended to furnish them with essential details of the training in which they will participate. *Consent forms* were provided in a printed format wherein participants were directed to submit the consent form to the researcher and to save a copy for their records. Participants also submitted a digital language background questionnaire aimed exclusively at gathering additional details about their language abilities. To ensure confidentiality, these completed questionnaires were securely stored on the Newcastle University researcher's OneDrive account.

3.8 Analysis

3.8.1 A generalised Linear Mixed Model (GLMM)

In this study, response accuracy is the dependent variable, while tests, groups, and vowels are the independent variables. The primary objective is to investigate the influence of tests (pre-, mid-, and post-tests, as well as the two generalisation tests), groups, and vowels on participants' responses, thus answering the study's primary research questions. A generalised Linear Mixed Model (GLMM) with a binomial family was utilised for the data analysis due to its exceptional proficiency in managing intricate data structures (Lammertyn et al., 2003; Agresti, 2012). Below is a justification for employing a GLMM with a binomial family to investigate the effects of predictors (tests, groups, vowels) on the outcome variable.

Binary Response Variable:

This study's outcome variable is binary, represented as 'correct' or 'incorrect'. Therefore, the binomial family is the appropriate choice for assessing the accuracy of these binary outcomes. This method models binary data by connecting the linear predictor to the response via a logistic function, ensuring the resultant probabilities lie between 0 and 1 (Agresti 2012).

Fixed and random effects:

GLMM incorporates both fixed and random effects, presenting a holistic approach that deepens the analysis of data, which in turn produces more reliable findings regarding the factors that influence response accuracy. Although fixed factors play a significant role in determining response accuracy, recognising random effects in the GLMM structure assists in pinpointing inherent variabilities. This further enhances the results' credibility and applicability by accounting for variability that cannot be explained by fixed factors alone (Zimmermann 1993).

Interactions:

With multiple predictors such as tests, vowels, and groups, interactions between these variables are possible. These interactions can be modelled by GLMMs, enabling an examination of whether the effect of one predictor varies based on the level of another predictor (Shang et al. 2018).

Chapter 4. Results

This chapter presents a comprehensive examination of the experimental results and data analysis, structured into five sections. The first section offers an overview of the analytical methods and the preliminary procedures employed to select the optimal model for the data. The second section includes the interpretation of the results. The third and fourth sections investigate the training and generalisation effects, concentrating on the outcomes of three perceptual tasks: identification (ID), auditory discrimination (AD), and category discrimination (CD). For each task, the study conducts the following analyses:

1. Comparing the tests (pre-, mid-, and post-tests, or generalisation tests in comparison to pre- and post-tests) across different groups.
2. Comparing the performance of three experimental groups across different tests: Group A (exposed to single L1 input), Group B (exposed to multiple L1 inputs), and Group C (exposed to multiple L1 inputs and a single L2 input).
3. Investigating the accuracy of vowels across different tests.

The final section concludes with a comprehensive summary of the findings.

4.1 Overview of analysis methods and model selection procedures

The analysis used a Generalised Linear Mixed Modelling (GLMM) approach with a binomial family to examine the overall effects of ***tests, groups, and vowels*** on the response accuracy, using the lme4 package in R (Bates et al., 2015).The goal was to identify the best-fitting model by applying a series of binomial logistic mixed effect models to the response data, through the glmer function. Different logistic mixed-effect models were created for three perceptual tasks (ID, AD, CD), with some accounting for interactions and others not (Al-Tamimi, 2022, 2023) (see table 4.1).

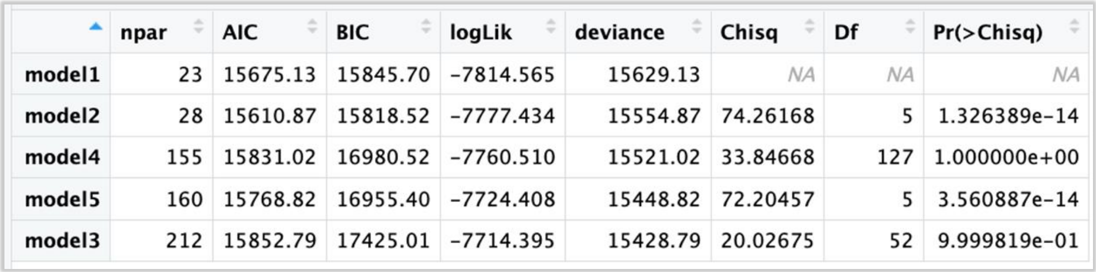
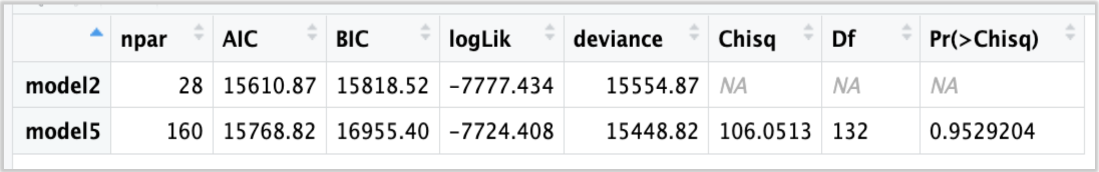
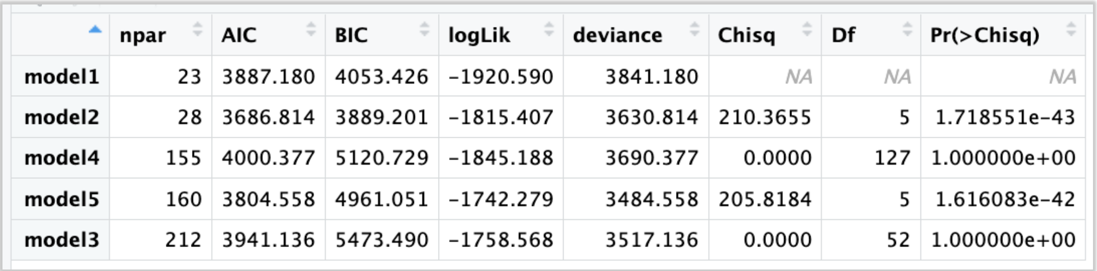
The analysis of three tasks (ID, AD, and CD) started with Model 1 as the baseline. It included test, group, and vowel as fixed factors based on the assumption that they would reveal key differences. It also accounts for random intercepts for participants and variability in word production across different tests and speakers. Model 2 increases complexity by adding a random slope for the test variable across participants, allowing for individual variation in test responses. Building on Model 2, Model 3 introduces an additional random slope for the interaction between the test and vowel variables, further accounting for individual differences. Model 4 includes interaction terms among the test, group, and vowel variables but does not incorporate random slopes, allowing examination of potential interaction effects without added complexity. In Model 5, both interaction terms and random slopes are added, providing the most comprehensive analysis by combining the complexities of Models 3 and 4. Beyond Model 5, more complex models failed to converge, marking the limit of analytical complexity. Therefore, model comparisons were conducted up to Model 5.

<table>
<tr><th>Tasks</th><th>List of models</th></tr>
<tr><td><b>ID</b></td><td>Model.1 <- dt2 %>% glmer(Results_ID ~ Test + Group + Vowel + (1|participant) + (1|Word_test_speaker), data = ., family = binomial, control = glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e5)))
Model.2 <- dt2 %>% glmer(Results_ID ~ Test + Group + Vowel + (test|participant) + (1| Word_test_speaker), data = ., family = binomial, control = glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e5)))
Model.3 <- dt2 %>% glmer(Results_ID ~ Test + Group + Vowel + (test + Vowels_ID|participant) + (1| Word_test_speaker), data = ., family = binomial, control = glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e5)))
Model.4 <- dt2 %>% glmer(Results_ID ~ Test * Group * Vowels + (1|participant) + (1| Word_test_speaker), data = ., family = binomial, control = glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e5)))
Model.5 <- dt2 %>% glmer(Results_ID ~ Test * Group * Vowels + (test|participant) + (1| Word_test_speaker), data = ., family = binomial, control = glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e5)))</td></tr>
</table>

<table>
<tr><td><b>AD</b></td><td>Model.1 <- dt2 %>% glmer(Results_AD ~ Test + Group + Vowel + (1|participant) + (1| Word_test_speaker), data = ., family = binomial, control = glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e5)))
Model.2 <- dt2 %>% glmer(Results_AD ~ Test + Group + Vowel + (test|participant) + (1| Word_test_speaker), data = ., family = binomial, control = glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e5)))
Model.3 <- dt2 %>% glmer(Results_AD ~ Test + Group + Vowel + (test + Vowels_ID|participant) + (1| Word_test_speaker), data = ., family = binomial, control = glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e5)))
Model.4 <- dt2 %>% glmer(Results_AD ~ Test * Group * Vowels + (1|participant) + (1| Word_test_speaker), data = ., family = binomial, control = glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e5)))
Model.5 <- dt2 %>% glmer(Results_AD ~ Test * Group * Vowels + (test|participant) + (1| Word_test_speaker), data = ., family = binomial, control = glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e5)))</td></tr>
<tr><td><b>CD</b></td><td>Model.1 <- dt2 %>% glmer(Results_CD ~ Test + Group + Vowel + (1|participant) + (1| Word_test_speaker), data = ., family = binomial, control = glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e5)))
Model.2 <- dt2 %>% glmer(Results_CD ~ Test + Group + Vowel + (test|participant) + (1| Word_test_speaker), data = ., family = binomial, control = glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e5)))
Model.3 <- dt2 %>% glmer(Results_CD ~ Test + Group + Vowel + (test + Vowels_ID|participant) + (1| Word_test_speaker), data = ., family = binomial, control = glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e5)))
Model.4 <- dt2 %>% glmer(Results_CD ~ Test * Group * Vowels + (1|participant) + (1| Word_test_speaker), data = ., family = binomial, control = glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e5)))
Model.5 <- dt2 %>% glmer(Results_CD ~ Test * Group * Vowels + (test|participant) + (1| Word_test_speaker), data = ., family = binomial, control = glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 2e5)))</td></tr>
</table>

Table 4.1 Five logistic mixed-effect models for the three perceptual tasks (ID, AD, CD).

Model comparisons were then conducted using the anova() function in R, which employs a likelihood ratio test to select the best-fitting model for the data. For each perceptual task (ID, AD, CD), Models 1, 2, 3, 4, and 5 were compared, considering the inclusion of pre-, mid-, and post-tests (see Table 4.2 below). An examination of the model comparisons for all tasks reveals that the chi-squared tests' p-values between Model 1 and Model 2 and between Model 1 and Model 5 are below 0.05. Therefore, it is feasible to reject the null hypotheses at a significance level of 0.05, which implies that Models 2 and 5 outperform Model 1. Subsequently, models 2 and 5 were compared. Based on the output of the model comparison, it can be observed that the p-value of the chi-squared test conducted between models 2 and 5 is more significant than 0.05. Therefore, at a significance level of 0.05, the null hypothesis cannot be rejected, suggesting that model 2 performs better than model 5. Similar outcomes were observed with other test combinations, including (generalisation test 1, pre-test, post-test) and (generalisation test 2, pre-test, post-test). Each combination consistently showed that Model 2 is the optimal model.

<table>
<tr><th>Task</th><th>Model Comparisons</th><th>Optimal Model</th></tr>
<tr><td>ID</td><td>

<table>
<tr><th></th><th>npar</th><th>AIC</th><th>BIC</th><th>logLik</th><th>deviance</th><th>Chisq</th><th>Df</th><th>Pr(>Chisq)</th></tr>
<tr><td>model1</td><td>23</td><td>15675.13</td><td>15845.70</td><td>-7814.565</td><td>15629.13</td><td>NA</td><td>NA</td><td>NA</td></tr>
<tr><td>model2</td><td>28</td><td>15610.87</td><td>15818.52</td><td>-7777.434</td><td>15554.87</td><td>74.26168</td><td>5</td><td>1.326389e-14</td></tr>
<tr><td>model4</td><td>155</td><td>15831.02</td><td>16980.52</td><td>-7760.510</td><td>15521.02</td><td>33.84668</td><td>127</td><td>1.000000e+00</td></tr>
<tr><td>model5</td><td>160</td><td>15768.82</td><td>16955.40</td><td>-7724.408</td><td>15448.82</td><td>72.20457</td><td>5</td><td>3.560887e-14</td></tr>
<tr><td>model3</td><td>212</td><td>15852.79</td><td>17425.01</td><td>-7714.395</td><td>15428.79</td><td>20.02675</td><td>52</td><td>9.999819e-01</td></tr>
</table>

<table>
<tr><th></th><th>npar</th><th>AIC</th><th>BIC</th><th>logLik</th><th>deviance</th><th>Chisq</th><th>Df</th><th>Pr(>Chisq)</th></tr>
<tr><td>model2</td><td>28</td><td>15610.87</td><td>15818.52</td><td>-7777.434</td><td>15554.87</td><td>NA</td><td>NA</td><td>NA</td></tr>
<tr><td>model5</td><td>160</td><td>15768.82</td><td>16955.40</td><td>-7724.408</td><td>15448.82</td><td>106.0513</td><td>132</td><td>0.9529204</td></tr>
</table>
</td><td>Model 2</td></tr>
<tr><td>AD</td><td>

<table>
<tr><th></th><th>npar</th><th>AIC</th><th>BIC</th><th>logLik</th><th>deviance</th><th>Chisq</th><th>Df</th><th>Pr(>Chisq)</th></tr>
<tr><td>model1</td><td>23</td><td>3887.180</td><td>4053.426</td><td>-1920.590</td><td>3841.180</td><td>NA</td><td>NA</td><td>NA</td></tr>
<tr><td>model2</td><td>28</td><td>3686.814</td><td>3889.201</td><td>-1815.407</td><td>3630.814</td><td>210.3655</td><td>5</td><td>1.718551e-43</td></tr>
<tr><td>model4</td><td>155</td><td>4000.377</td><td>5120.729</td><td>-1845.188</td><td>3690.377</td><td>0.0000</td><td>127</td><td>1.000000e+00</td></tr>
<tr><td>model5</td><td>160</td><td>3804.558</td><td>4961.051</td><td>-1742.279</td><td>3484.558</td><td>205.8184</td><td>5</td><td>1.616083e-42</td></tr>
<tr><td>model3</td><td>212</td><td>3941.136</td><td>5473.490</td><td>-1758.568</td><td>3517.136</td><td>0.0000</td><td>52</td><td>1.000000e+00</td></tr>
</table>
</td><td>Model 2</td></tr>
</table>

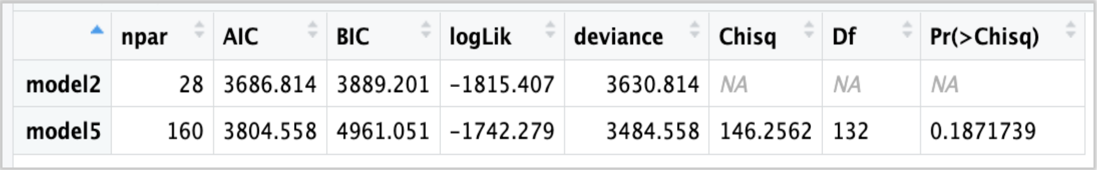
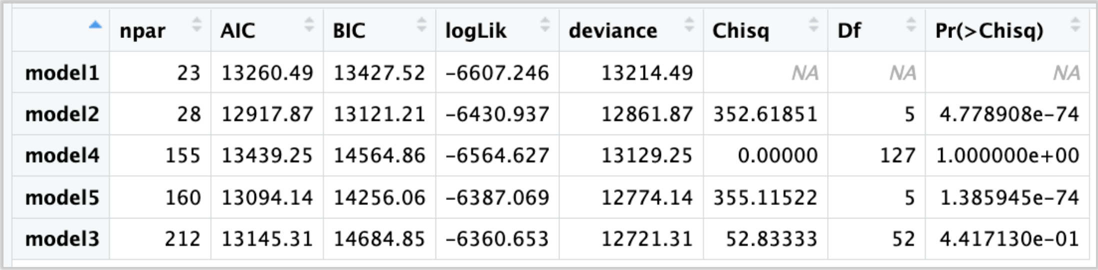
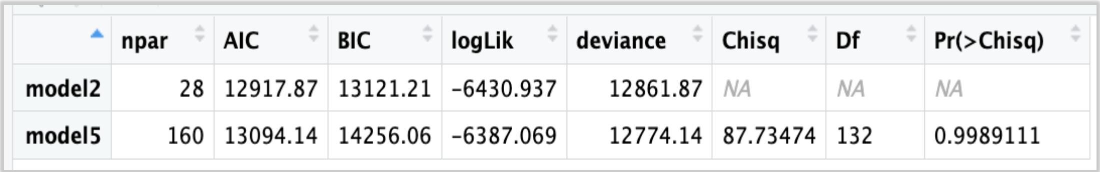
<table>
<tr>
<td></td>
<td>

<table>
<tr><th></th><th>npar</th><th>AIC</th><th>BIC</th><th>logLik</th><th>deviance</th><th>Chisq</th><th>Df</th><th>Pr(>Chisq)</th></tr>
<tr><td>model2</td><td>28</td><td>3686.814</td><td>3889.201</td><td>-1815.407</td><td>3630.814</td><td>NA</td><td>NA</td><td>NA</td></tr>
<tr><td>model5</td><td>160</td><td>3804.558</td><td>4961.051</td><td>-1742.279</td><td>3484.558</td><td>146.2562</td><td>132</td><td>0.1871739</td></tr>
</table>
</td>
<td></td>
</tr>
<tr>
<td><b>CD</b></td>
<td>

<table>
<tr><th></th><th>npar</th><th>AIC</th><th>BIC</th><th>logLik</th><th>deviance</th><th>Chisq</th><th>Df</th><th>Pr(>Chisq)</th></tr>
<tr><td>model1</td><td>23</td><td>13260.49</td><td>13427.52</td><td>-6607.246</td><td>13214.49</td><td>NA</td><td>NA</td><td>NA</td></tr>
<tr><td>model2</td><td>28</td><td>12917.87</td><td>13121.21</td><td>-6430.937</td><td>12861.87</td><td>352.61851</td><td>5</td><td>4.778908e-74</td></tr>
<tr><td>model4</td><td>155</td><td>13439.25</td><td>14564.86</td><td>-6564.627</td><td>13129.25</td><td>0.00000</td><td>127</td><td>1.000000e+00</td></tr>
<tr><td>model5</td><td>160</td><td>13094.14</td><td>14256.06</td><td>-6387.069</td><td>12774.14</td><td>355.11522</td><td>5</td><td>1.385945e-74</td></tr>
<tr><td>model3</td><td>212</td><td>13145.31</td><td>14684.85</td><td>-6360.653</td><td>12721.31</td><td>52.83333</td><td>52</td><td>4.417130e-01</td></tr>
</table>

<table>
<tr><th></th><th>npar</th><th>AIC</th><th>BIC</th><th>logLik</th><th>deviance</th><th>Chisq</th><th>Df</th><th>Pr(>Chisq)</th></tr>
<tr><td>model2</td><td>28</td><td>12917.87</td><td>13121.21</td><td>-6430.937</td><td>12861.87</td><td>NA</td><td>NA</td><td>NA</td></tr>
<tr><td>model5</td><td>160</td><td>13094.14</td><td>14256.06</td><td>-6387.069</td><td>12774.14</td><td>87.73474</td><td>132</td><td>0.9989111</td></tr>
</table>
</td>
<td>Model 2</td>
</tr>
</table>

Table 4.2 Outputs of model comparisons (Models 1, 2, 3, 4, and 5) for the tasks (ID, AD, CD), considering the inclusion of pre-, mid-, and post-tests

As demonstrated above, Model 2 (the model without interaction) emerged as the best fit for the data based on Log-Likelihood Chi-Squared tests. This model accounted for the individual influences of test, group, and vowel as fixed effects. The random structure was specified as (Test | Participant & 1 | Word_Test_Speaker), with the former representing a random slope and the latter representing a random intercept. See Table 4.3 for the formula of the best-fit models for ID, AD, and CD tasks and Table 4.4 for explanations of all factors in the model, including predictors, dependent variables, and random effects.

Tasks	The best-fitting model
ID	Results_ID ~ Test + Group + Vowels + (Test | Participant) + (1 | Word_Test_Speaker)
AD	Results_AD ~ Test + Group + Vowels + (Test | Participant) + (1 | Word_Test_Speaker)
CD	Results_CD ~ Test + Group + Vowels + (Test | Participant) + (1 | Word_Test_Speaker)

Table 4.3 The formula of the best-fit model for ID, AD, and CD tasks

Independent Variables (Predictors, Fixed Effects)	Categories (levels)	Dependent Variables	Categories (levels)	Notes
Test	**Pre**-, Mid-, Post **Pre**-, Post-, Gen1 **Pre**-, Post-, Gen2	Results_ID Results_AD Results_CD	**Incorrect**, Correct	Bold text indicates the reference category.
Group	**A**, B, C			
Vowel	**/æ/**,/aɪ/, /aʊ/, /ɑː/, /ɒ/, /e/, /eɪ/, /əʊ/, /ɛə/, /ɜː/, /iː/, /ɪ/, /ɔː/, /ɔɪ/, /uː/, /ʊ/, /ʌ/			
Random Effects	**Explanation**			
Test | Participant	The term "(Test | Participant)" represents a random slope for the "test" variable by participant, accounting for individual differences in responses to the tests. This allows the relationship between the response variable and the test predictor to vary by participant.			
1|Word_Test_Speaker	The term "1 | Word_Test_Speaker" represents a random intercept. This intercept captures the variability in word production across different tests and speakers. Each speaker produces different words for each test, ensuring that the same words are not repeated across tests or by the same speaker within a test. For example, 'coach_Pre-test_SSBE speaker', 'coach_Post-test_SSBE speaker', and 'stone_gen1-test_Saudi speaker'. Furthermore, listeners are exposed to multiple accents from various speakers in each training set, and the model accounts for this diversity. Initially, random slopes were attempted but were unsuccessful, leading to the use of a nested model to address the different words produced by each speaker for each test.			

Table 4.4 Explanation of all factors included in the optimal model

4.2 Results discussion

To interpret the results, the explanation of each fixed factor (test, group, and vowel) is based on the concept of probability. Probability ranges from 0 to 1, or from 0% to 100%. There is a direct relationship between odds ratio and probability: odds ratio is greater than 1 if the probability is above 0.5 (50%), and less than 1 if it is below 0.5. Odds ratio is exactly 1 when the probability is 0.5 (50%) (Nahhas, 2024). In logistic regression, a log-odds value of 0 corresponds to an odds ratio of 1, equating to a 50% probability. This pivotal point signifies no difference between the likelihood of an event occurring or not occurring. Consequently, any log-odds greater than 0 indicates an increased likelihood of the event occurring, with probabilities exceeding 50%. Conversely, log-odds less than 0 indicate a decreased likelihood, with probabilities falling below 50% (Introduction to SAS. UCLA: Statistical Consulting Group, 2024).

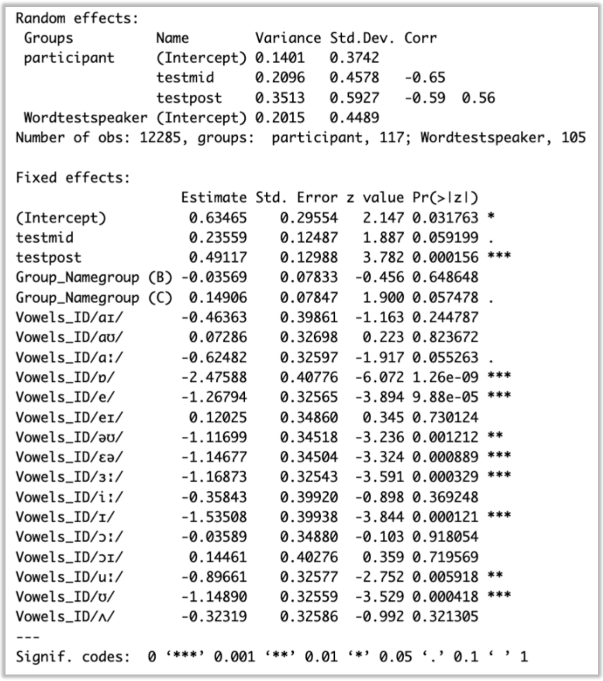
The 50% threshold is a widely accepted benchmark in binary classification tasks due to its foundation in statistical principles. In this study, the dependent variable "response" can have two outcomes: "correct" or "incorrect." When responses are evenly split, with 50% correct and 50% incorrect, the threshold probability is 50%, indicating an equal likelihood of either outcome. Thus, the 50% threshold serves as a natural midpoint for interpreting probabilities and understanding performance levels. It is not a significant threshold but rather a reference point for distinguishing between higher and lower probabilities of correct responses, distinct from the p-value (probability of type I error).

To clarify, the 50% threshold used in this study represents the total number of correct responses divided by the total number of responses, rather than a comparison of correct to incorrect responses. It does not imply that scores below 50% are incorrect and those above are correct. Instead, it serves as a benchmark for assessing the performance of groups across tests, tests across groups, and vowels across tests. Scores above 50% (e.g., 53%, 68%, 75%) indicate a higher probability of correct responses, while scores below 50% (e.g., 14%, 33%, 46%) indicate a lower probability of correct responses. In short, the 50% threshold serves as an indicator of the likelihood of correct responses, marking an increase or decrease in this probability.

4.3 The effects of training

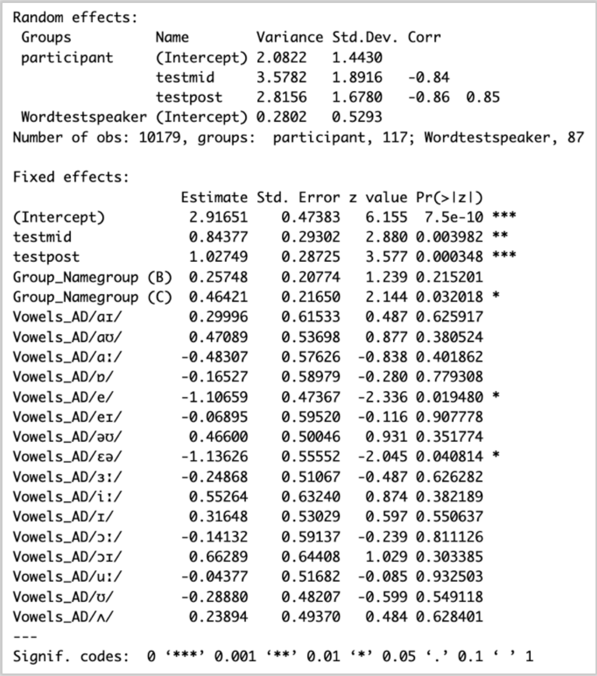
This section reports the results of the pre-, mid-, and post-tests for ID, AD, and CD tasks. The table below provides a comprehensive summary of the primary fixed effects (test, group, and vowel) along with the random effects for each task.

ID



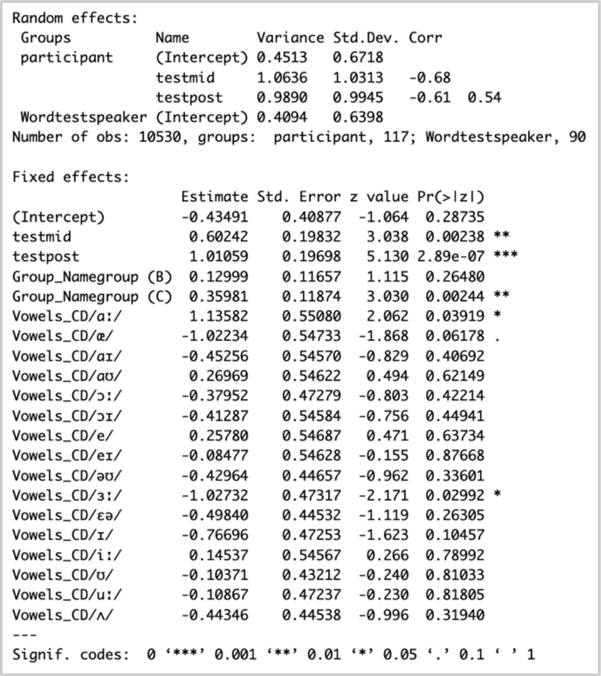
```
Random effects:
 Groups          Name        Variance Std.Dev. Corr
 participant     (Intercept) 0.1401   0.3742
                 testmid     0.2096   0.4578   -0.65
                 testpost    0.3513   0.5927   -0.59  0.56
 Wordtestspeaker (Intercept) 0.2015   0.4489
Number of obs: 12285, groups:  participant, 117; Wordtestspeaker, 105

Fixed effects:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          0.63465    0.29554   2.147 0.031763 *
testmid              0.23559    0.12487   1.887 0.059199 .
testpost             0.49117    0.12988   3.782 0.000156 ***
Group_Namegroup (B) -0.03569    0.07833  -0.456 0.648648
Group_Namegroup (C)  0.14906    0.07847   1.900 0.057478 .
Vowels_ID/aɪ/       -0.46363    0.39861  -1.163 0.244787
Vowels_ID/aʊ/        0.07286    0.32698   0.223 0.823672
Vowels_ID/ɑː/       -0.62482    0.32597  -1.917 0.055263 .
Vowels_ID/ɒ/        -2.47588    0.40776  -6.072 1.26e-09 ***
Vowels_ID/e/        -1.26794    0.32565  -3.894 9.88e-05 ***
Vowels_ID/eɪ/        0.12025    0.34860   0.345 0.730124
Vowels_ID/əʊ/       -1.11699    0.34518  -3.236 0.001212 **
Vowels_ID/ɛə/       -1.14677    0.34504  -3.324 0.000889 ***
Vowels_ID/ɜː/       -1.16873    0.32543  -3.591 0.000329 ***
Vowels_ID/iː/       -0.35843    0.39920  -0.898 0.369248
Vowels_ID/ɪ/        -1.53508    0.39938  -3.844 0.000121 ***
Vowels_ID/ɔː/       -0.03589    0.34880  -0.103 0.918054
Vowels_ID/ɔɪ/        0.14461    0.40276   0.359 0.719569
Vowels_ID/uː/       -0.89661    0.32577  -2.752 0.005918 **
Vowels_ID/ʊ/        -1.14890    0.32559  -3.529 0.000418 ***
Vowels_ID/ʌ/        -0.32319    0.32586  -0.992 0.321305
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

AD



```
Random effects:
 Groups          Name        Variance Std.Dev. Corr
 participant     (Intercept) 2.0822   1.4430
                 testmid     3.5782   1.8916   -0.84
                 testpost    2.8156   1.6780   -0.86  0.85
 Wordtestspeaker (Intercept) 0.2802   0.5293
Number of obs: 10179, groups:  participant, 117; Wordtestspeaker, 87

Fixed effects:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          2.91651    0.47383   6.155  7.5e-10 ***
testmid              0.84377    0.29302   2.880 0.003982 **
testpost             1.02749    0.28725   3.577 0.000348 ***
Group_Namegroup (B)  0.25748    0.20774   1.239 0.215201
Group_Namegroup (C)  0.46421    0.21650   2.144 0.032018 *
Vowels_AD/aɪ/        0.29996    0.61533   0.487 0.625917
Vowels_AD/aʊ/        0.47089    0.53698   0.877 0.380524
Vowels_AD/ɑː/       -0.48307    0.57626  -0.838 0.401862
Vowels_AD/ɒ/        -0.16527    0.58979  -0.280 0.779308
Vowels_AD/e/        -1.10659    0.47367  -2.336 0.019480 *
Vowels_AD/eɪ/       -0.06895    0.59520  -0.116 0.907778
Vowels_AD/əʊ/        0.46600    0.50046   0.931 0.351774
Vowels_AD/ɛə/       -1.13626    0.55552  -2.045 0.040814 *
Vowels_AD/ɜː/       -0.24868    0.51067  -0.487 0.626282
Vowels_AD/iː/        0.55264    0.63240   0.874 0.382189
Vowels_AD/ɪ/         0.31648    0.53029   0.597 0.550637
Vowels_AD/ɔː/       -0.14132    0.59137  -0.239 0.811126
Vowels_AD/ɔɪ/        0.66289    0.64408   1.029 0.303385
Vowels_AD/uː/       -0.04377    0.51682  -0.085 0.932503
Vowels_AD/ʊ/        -0.28880    0.48207  -0.599 0.549118
Vowels_AD/ʌ/         0.23894    0.49370   0.484 0.628401
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

CD



```
Random effects:
 Groups          Name        Variance Std.Dev. Corr
 participant     (Intercept) 0.4513   0.6718
                 testmid     1.0636   1.0313   -0.68
                 testpost    0.9890   0.9945   -0.61  0.54
 Wordtestspeaker (Intercept) 0.4094   0.6398
Number of obs: 10530, groups:  participant, 117; Wordtestspeaker, 90

Fixed effects:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         -0.43491    0.40877  -1.064  0.28735
testmid              0.60242    0.19832   3.038  0.00238 **
testpost             1.01059    0.19698   5.130 2.89e-07 ***
Group_Namegroup (B)  0.12999    0.11657   1.115  0.26480
Group_Namegroup (C)  0.35981    0.11874   3.030  0.00244 **
Vowels_CD/ɑː/        1.13582    0.55080   2.062  0.03919 *
Vowels_CD/æ/        -1.02234    0.54733  -1.868  0.06178 .
Vowels_CD/aɪ/       -0.45256    0.54570  -0.829  0.40692
Vowels_CD/aʊ/        0.26969    0.54622   0.494  0.62149
Vowels_CD/ɔː/       -0.37952    0.47279  -0.803  0.42214
Vowels_CD/ɔɪ/       -0.41287    0.54584  -0.756  0.44941
Vowels_CD/e/         0.25780    0.54687   0.471  0.63734
Vowels_CD/eɪ/       -0.08477    0.54628  -0.155  0.87668
Vowels_CD/əʊ/       -0.42964    0.44657  -0.962  0.33601
Vowels_CD/ɜː/       -1.02732    0.47317  -2.171  0.02992 *
Vowels_CD/ɛə/       -0.49840    0.44532  -1.119  0.26305
Vowels_CD/ɪ/        -0.76696    0.47253  -1.623  0.10457
Vowels_CD/iː/        0.14537    0.54567   0.266  0.78992
Vowels_CD/ʊ/        -0.10371    0.43212  -0.240  0.81033
Vowels_CD/uː/       -0.10867    0.47237  -0.230  0.81805
Vowels_CD/ʌ/        -0.44346    0.44538  -0.996  0.31940
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 4.5 Overview of model outputs, detailing fixed and random effects for ID, AD, and CD tasks (training effects: pre-, mid-, and post-tests).

The following subsections primarily examine the fixed effects (test, group, and vowel) for the ID, AD, and CD tasks. The approaches utitlised for analysing each fixed effect are detailed below:

1. A comprehensive explanation of the statistical summary for the logistic mixed-effects regression model.
2. Pairwise comparison tests (Wald test) conducted using the Estimated Marginal Means (EMMeans) package (without interaction), accompanied by model-generated visualisations. Since the raw data visualisations led to similar conclusions, these figures are included in the appendix.
3. Multiple comparison t-tests of predicted values based on the raw data (with interaction).

4.3.1 Vowel Identification

4.3.1.1 Tests

The plogis() function was applied to provide proportional responses for the coefficients of a mixed-effect model, facilitating the detection of individual test effects on response accuracy for the ID task using the pre-test as a baseline for comparison against the mid- and post-tests. The findings demonstrated a significant increase in the probability of correct responses for the post-test (62%, p < 0.0001), while there was a tendency for the mid-test to be significant (56%, p = 0.059).

The EMMeans package was utilised to perform a comprehensive pairwise comparison analysis of tests (Wald test) to thoroughly examine potential differences (Table 4.6). As illustrated, the odds of accurate responses in the pre-test were less than 1, specifically 0.612 times compared to that of the post-test. The very low p-value (p < 0.001) implies a significant improvement in the accuracy of responses in the post-test compared to the pre-test, with the odds ratio being approximately 1.63 (1/0.612). Conversely, the odds ratios between pre- and mid-tests and mid- and post-tests were 0.790 and 0.774, respectively. Both comparisons yield a p-value of 0.0592, indicating a tendency towards significant differences between these contrasts. In other words, the mid-test demonstrated a modest improvement in accuracy relative to the pre-test, whereas the post-test demonstrated a similar increase in accuracy relative to the mid-test.

Contrast	***Odds.ratio***	***SE***	***z.ratio***	***p.value***
pre / mid	0.790	0.0987	-1.887	0.0592
pre / post	0.612	0.0795	-3.782	0.0005
mid / post	0.774	0.0983	-2.013	0.0592

Table 4.6 The main effect of the test phase (pre-, mid-, and post- test) on the ID task, excluding interactions. The mid-point represents an odds ratio of 1.

The plot provided below depicts the EMMeans of test accuracy among groups A, B, and C. These EMMeans relate to the probability of correct answers for pre-, mid, and post-tests within each group. As shown, there was a progressive enhancement in performance across all groups over time, starting from the pre-test to the mid-test and concluding with the post-test. The confidence intervals for the average predicted probabilities of tests across the groups appear to be stable. This consistency in the confidence intervals (CIs) implies a steady level of accuracy in the estimations of the mean test scores.

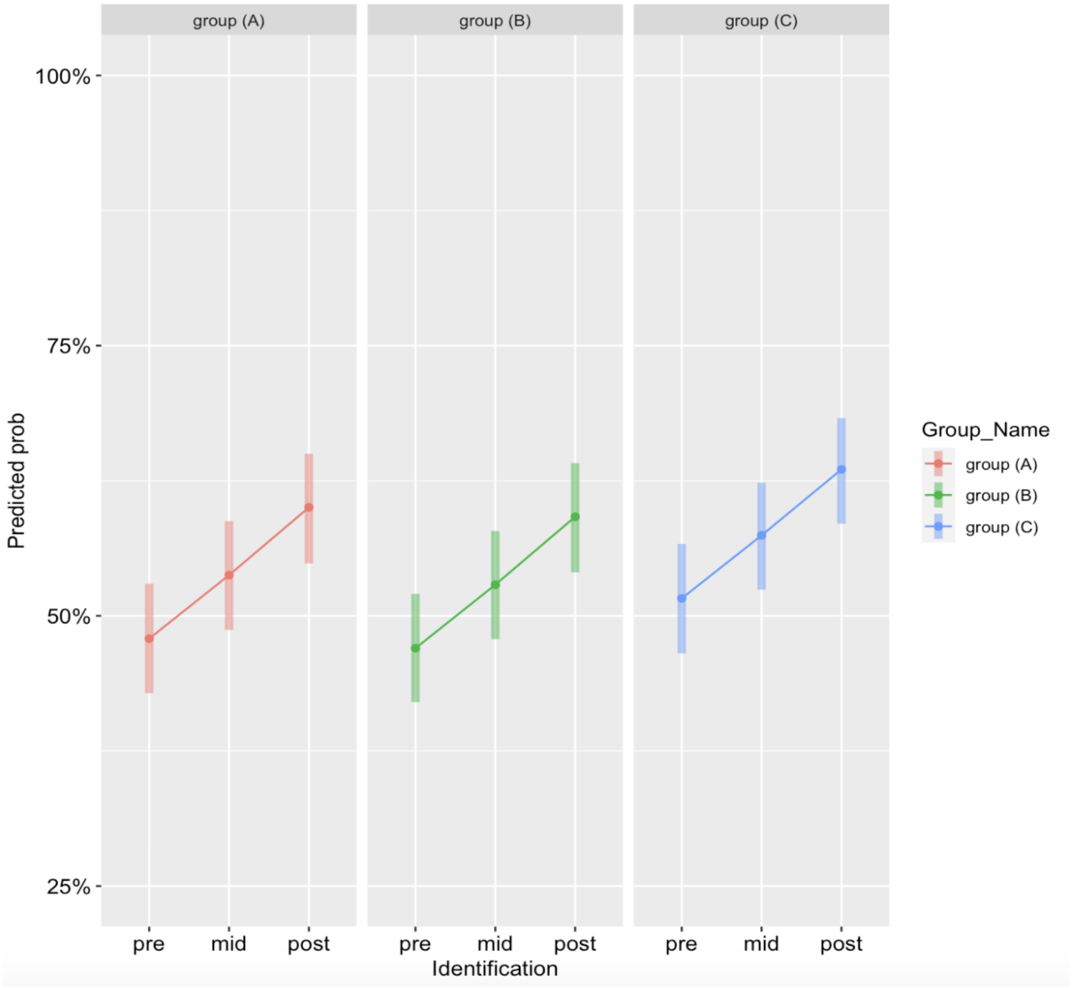


Figure 4.1 The EMMeans of pre-, mid-, and post-tests across groups for the ID task (50% threshold)

The interactions between the test contrasts of the ID task (pre- vs. mid, pre- vs. post- and mid vs. post-) across experimental groups were determined using the multiple comparison t-test (see Table 4.7). The performance of the three groups on the post-test was significantly greater than their performance on the pre-test (by 11%, $p < 0.0001$) and the mid-test (by 6%, $p < 0.0001$). Similarly, the mid-test performance significantly outperformed the pre-test performance for all three groups (by 5%, $p < 0.0001$).

Groups	***Contrast (test)***	***Mean Difference***	***p-value***
A	pre – mid	-5%	0.000
	pre – post	-11%	0.000
	mid – post	-6%	0.000
B	pre – mid	-5%	0.000
	pre – post	-11%	0.000
	mid – post	-6%	0.000
C	pre – mid	-5%	0.000
	pre – post	-11%	0.000
	mid – post	-6%	0.000

Table 4.7 Interactions of pre-, mid-, and post-tests across groups for the ID task[52].

4.3.1.2 Groups

The analysis of mixed-effect model coefficients highlights the performance of the experimental groups across the pre-, mid-, and post-tests for the ID task. The results revealed a tendency in the probability of response accuracy to be significant for group C (54%, $p = 0.057$) compared to the reference category (group A), while no significant difference was observed for group B (49%, $p > 0.05$).

[52] The mean difference is computed as the proportion of responses at the pre-test – proportion of responses at the mid-test or post-test. When the proportion at the mid-test (or post-test) is higher than that of the pre-test, this yields a negative mean difference. The same principle applies to *mid – post*

Table 4.8 provides a side-by-side examination of the groups. The odds of accuracy responses were roughly similar when comparing group A with B (odds ratio = 1.036, *p* > 0.05). This suggests no significant discrepancy in the performance levels of these two groups. However, the odds of providing accurate responses were lower for both group A (odds ratio = 0.862) and group B (odds ratio = 0.831) when each was compared to group C. The p values of 0.086 and 0.0532 are above the traditional statistical significance threshold of 0.05 but still hint at a possible trend: group C tends to have a better performance or a higher likelihood of providing correct answers than groups A and B.

Contrast	**Odds.ratio**	**SE**	**z.ratio**	**p.value**
group (A) / group (B)	1.036	0.0812	0.456	0.6486
group (A) / group (C)	0.862	0.0676	-1.900	0.0862
group (B) / group (C)	0.831	0.0648	-2.371	0.0532

Table 4.8 Main effect of groups (A, B, C) in the ID task across the pre-, mid-, and post-tests, excluding interactions. The odds ratio midpoint is 1.

Figure 4.2 presents EMMeans for groups A, B, and C across the pre-, mid-, and post-tests. These EMMeans represent the probabilities of correct responses for each group in the respective tests. The CIs for the average predicted probabilities of the groups across all tests show steady consistency. Such sustained regularity in the confidence intervals implies a constant level of accuracy in the estimations of the average scores for each group. Upon a general observation of the plot, it is evident that all three groups demonstrated improvement over time. In the pre-test, group A had an estimated probability of 48%, group B had a probability of 47%, and group C had a probability of 52%. These values reflect the initial performance levels of each group prior to conducting the training. The probabilities of accurate responses continued to rise in the mid- and post-tests for all groups, with group C, exhibiting a slightly higher probability of accurate responses than groups A and B. The conclusion can be drawn that groups A, B, and C demonstrate comparable performance for the

ID task. The higher level of accuracy observed in group C's performance during the mid-and post-tests does not indicate that they are outperforming groups A and B. Rather, it should be noted that group C also demonstrated better performance in the pre-test.

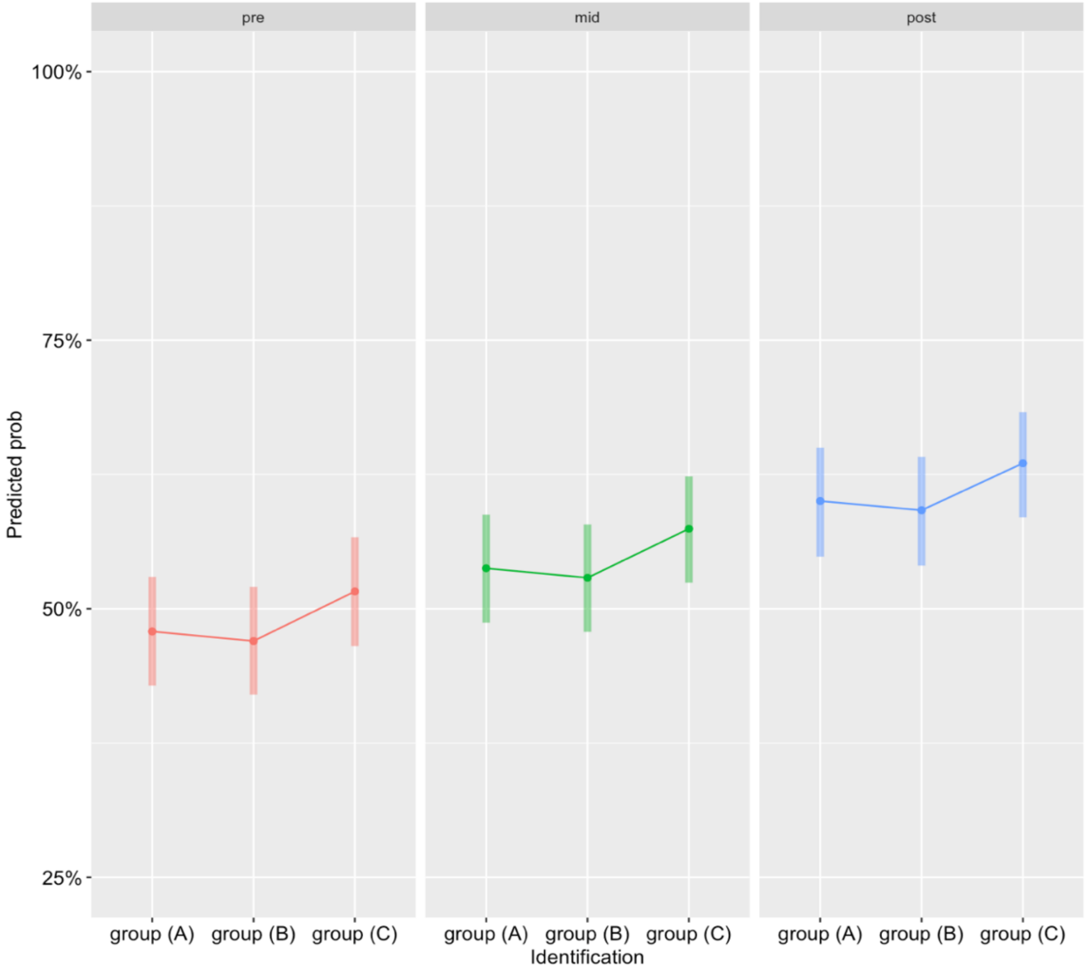


Figure 4.2 The EMMeans of groups (A, B, C) across pre-, mid-, and post-tests for the ID task (threshold set at 50%).

The multiple comparison t-test was performed to determine the *interactions* of groups (A vs B, A vs C, and B vs C) across the tests for the ID task (see Table 4.9 below). During the pre-, mid-, and post-tests, there were no substantial variations between group A and group B. There was, however, a statistically significant difference between groups A and C and between groups B and C (*p* < 0.0001). Group C outperformed groups A and B by 3 to 5%. Again, as group C performed better on the pre-test than groups A and B, their higher scores on the mid-test and post-test do not indicate improved performance. Yet, these results suggest that all three groups performed similarly.

Test	*Contrast (groups)*	*Mean Difference*	*p-value*
Pre-test	A – B	1%	0.12
	A – C	-4%	0.000
	B – C	-5%	0.000
Mid-test	A – B	1%	0.12
	A – C	-3%	0.000
	B – C	-3%	0.000
Post-test	A – B	1%	0.12
	A – C	-4%	0.000
	B – C	-4%	0.000

Table 4.9 Interactions of groups across tests pre-, mid-, and post-tests for the ID task[53].

4.3.1.3 Vowels

The multiple comparison t-test for the ID task revealed the interactions of tests (pre-, mid-, and post-) across vowels (See Table G.I in Appendix F). Relative to the pre-test, there was a marked improvement in vowel accuracy in the post-test, with improvements ranging from 7% to 12% (p < 0.0001), and in the mid-test, with improvements ranging from 3% to 6% (p < 0.0001). Moreover, the mid-test results demonstrated a significant enhancement over the post-test results, with vowel performance rising by 4% to 6% (p<0.0001).

Table 4.10 shows the individual effects of 17 vowels on response accuracy across the three tests without interactions. Probabilities surpassing the 50% threshold were colour-coded as green, while those falling below the threshold were represented in red. The colour gradient serves as an indication of improvement. As shown, the

[53] The mean difference is calculated by subtracting the proportion of responses of group C from the proportion of responses of groups A and B across the tests (pre, mid, and post). Since the proportion for group C is greater than those groups A and B, the mean difference is shown as negative

vowels /ɔɪ/, /eɪ/, /aʊ/, /æ/ and /ɔː/ had the highest probability of correct responses during the pre-test, with individual rates of 69%, 69%, 68%, 66%, and 65% respectively. This was followed by a 5% increase in the probability of accurate answers for these vowels in the mid-test compared to the pre-test results. The post-test saw further improvement compared to the pre-test, with the accuracy rate escalating by 9% for /ɔɪ/ and /eɪ/ and 10% for /aʊ/, /æ/ and /ɔː/. In addition, compared to the mid-test, the accuracy of these vowels in the post-test increased by 5%. In a parallel manner, the vowels /ʌ/, /iː/, /aɪ/, and /ɑː/ exhibited high probabilities of correct responses during the pre-test, achieving rates of 59%, 58%, 55%, and 51% respectively. In comparison to the pre-test results, the vowel accuracy during the post-tests showed consistent improvements over time, resulting in a 12% rise. A notable increase of 6% in the performance of the vowels was observed at the mid-test in comparison to the pre-test, and a similar improvement was also observed during the post-test in contrast to the mid-test.

On the other hand, the lowest probability of response accuracy was observed for the vowel /ɒ/ across the three tests. The likelihood for correct responses was merely 14% in the pre-test, increased marginally to 17% in the mid-test, and further to 21% in the post-test. These findings suggest that participants experienced difficulty in perceiving the /ɒ/ vowel; however, improvement was noted over time. In particular, there was a 3% increase in performance from the pre-test to the mid-test, a 7% increase from the pre-test to the post-test, and a 4% increase from the mid-test to the post-test. Moreover, the vowels /e/, /ɪ/, /uː/, /ʊ/, /ɜː/ and /ɛə/ had adverse effects on response accuracy during the pre-test (all scored below 50%). However, as time progressed, they demonstrated higher probabilities of correct responses. There was a 6% increase in the accuracy of the vowels /e/, /uː/, /ʊ/, /ɜː/, and /ɛə/ in the mid-test when compared to the pre-test. A comparable improvement (6%) was observed in the post-test when compared to the mid-test. The post-test results showed a 12% improvement in the performance of these vowels as compared to the pre-test. In regard to the vowel /ɪ/, there was a 5% improvement from the pre-test to the mid-test, an 11% increase from the pre-test to the post-test, and a 6% improvement from the mid-test to the post-test.

Vowels	***Pre-test***	***Mid-test***	***Post-test***
/ɔɪ/	69.4%	74.1%	78.7%
/eɪ/	68.8%	73.7%	78.3%
/aʊ/	67.8%	72.7%	77.5%
/æ/	66.2%	71.3%	76.2%
/ɔː/	65.4%	70.5%	75.5%
/ʌ/	58.6%	64.2%	69.9%
/iː/	57.8%	63.4%	69.1%
/aɪ/	55.2%	60.9%	66.8%
/ɑː/	51.2%	57.0%	63.2%
/uː/	44.4%	50.3%	56.6%
/əʊ/	39.1%	44.8%	51.2%
/ʊ/	38.3%	44.0%	50.4%
/ɛə/	38.4%	44.1%	50.4%
/ɜː/	37.8%	43.5%	49.9%
/e/	35.5%	41.1%	47.4%
/ɪ/	29.7%	34.8%	40.8%
/ɒ/	14.1%	17.3%	21.2%

Table 4.10 The EMMeans of vowels across pre-, mid-, and post-tests for the ID task[54].

The plots below visually depict the EMMeans of vowels across tests, along with their corresponding CIs. Two plots are generated: one for the vowels that were deemed easy, scoring above the 50% threshold (Figure 4.3), and another for the problematic vowels, scoring below the 50% threshold (Figure 4.4). On examining the **'easy'** vowels in Figure 4.3, specific patterns emerge. The vowels /ʌ/, /ɔː/, /eɪ/, /ɑː/, and /aʊ/ exhibit smaller CIs, signifying less variability around their average predicted

[54] The correct response probabilities are represented by the following colours: (dark green) for high, (medium green) for moderate, and (pale green) for low. Incorrect response probabilities use (dark red) to indicate high, (medium red) to represent indicate, and (pale red) to indicate low.

probabilities. This suggests that learners show greater consistency in their perception of these vowels, which might point to higher confidence levels. On the other hand, the vowels /ɔɪ/, /iː/, /aɪ/, and /æ/ are observed to display wider CIs, pointing to a larger variability around their average predictive probabilities. The wider range in learners' perception of these vowels indicates individual variations in their responses.

Moving the focus to the examination of the **'difficult'** vowels depicted in Figure 4.4, it is observed that most of the vowels (/ɒ/, /e/, /ɜː/, /ɛə/, /əʊ/, /ʊ/, and /uː/) — excluding /ɪ/ — exhibit narrower CIs, indicating less fluctuation around their average predicted probabilities. Given that these vowels received pre-test scores below the threshold of 50%, it is apparent that learners have difficulty accurately perceiving these sounds. However, with the passage of time, even these vowels, which were formerly regarded as more difficult, demonstrated a continuing improvement.

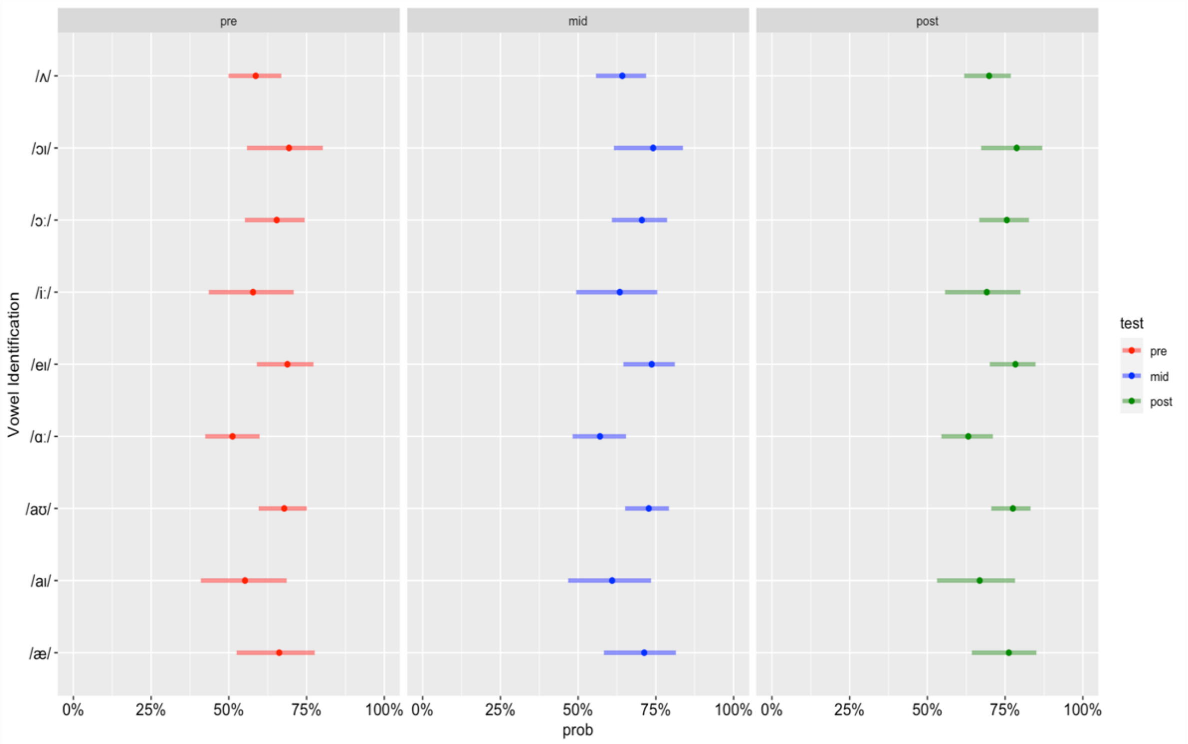


Figure 4.3 Vowels achieving scores above 50% chance of correct responses across the pre-, mid- and post-tests for the ID task (50% threshold).

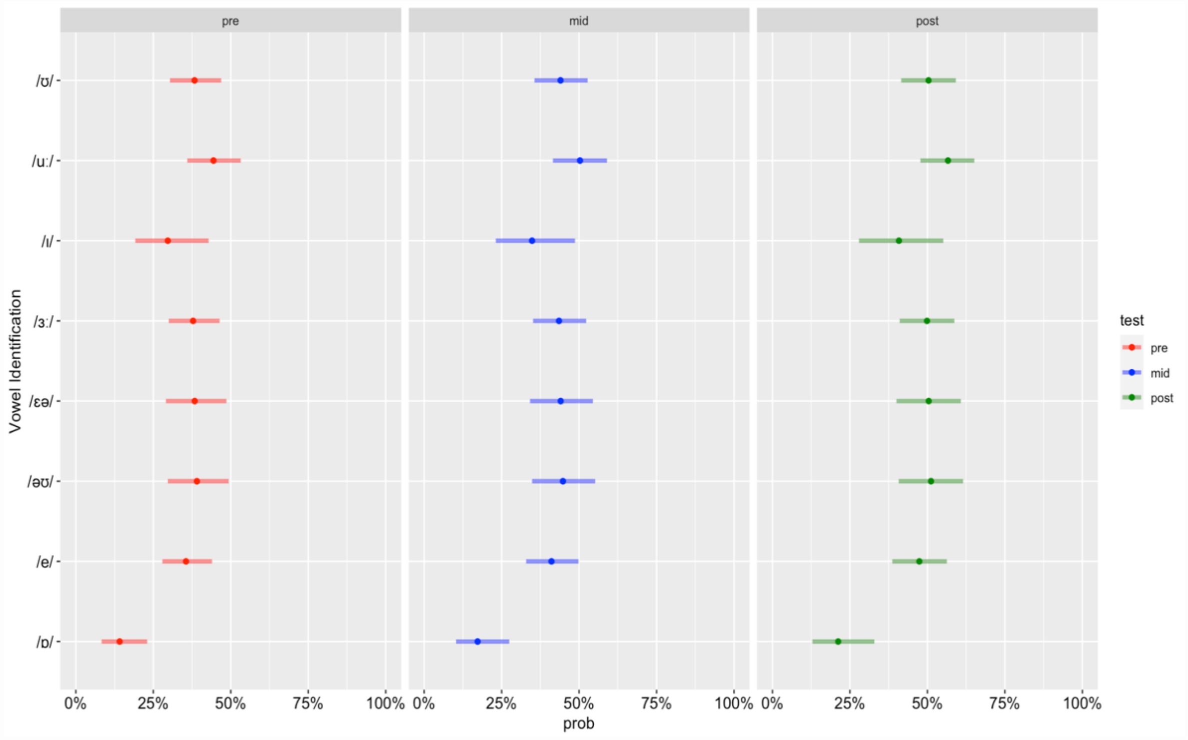


Figure 4.4 Vowels achieving scores below a 50% chance of correct responses across the pre-, mid-, and post-tests for the ID task (50% threshold).

In brief, during the vowel identification task, the vowels /ɒ/, /e/, /ɪ/, /ɜː/, /ɛə/, /əʊ/, /ʊ/, and /uː/ presented perceptual challenges for the participants, as the scores for these vowels fell beneath the 50% probability threshold for the ID task. Nevertheless, an upward trend was observed in the mid- and post-tests, indicating that the training facilitated the improvement of these difficult vowels over time. On the other hand, the vowels /ɔɪ/, /eɪ/, /aʊ/, /æ/, /ɔː/, /ʌ/, /iː/, /aɪ/, and /ɑː/ were easily perceived before the training, as they all exceeded the probability threshold of 50%. Over time, the easy vowels also showed gradual improvement.

4.3.1.4 Summary

The below points summarise the findings of the vowel identification task:

- A consistent increase in performance was noticed over time for all groups, from the initial pre-test through the mid-test and ultimately to the post-test.

- Across the pre-, mid-, and post-test stages, the performance of Groups A, B, and C remains relatively similar. This indicates the success of training, regardless of the diversity of inputs, whether the training involves a single L1 variety, multiple L1 varieties, or a combination of multiple L1 and one L2 variety.

- Training incorporating variations of accents proved to be non-problematic for learners with relatively low proficiency levels, making it highly suitable for application within foreign language classroom environments.

- During the identification task, the vowels /ɔɪ/, /eɪ/, /aʊ/, /æ/, /ɔː/, /ʌ/, /iː/, /aɪ/, and /ɑː/ were easily recognised, achieving accuracy rates over 50% even in the pre-test. However, the vowels /ɒ/, /e/, /ɪ/, /ɜː/, /ɛə/, /əʊ/, /ʊ/, and /uː/ were difficult to discern at the pre-test, as their accuracy rates were found to be below 50%.

- Both easy and challenging vowels demonstrated progress over time, with the post-test—after 16 training sessions—showing more improvement than the mid-test after 8 sessions. This suggests that in terms of vowel training, an extended duration of learning is advantageous for learners, even though some vowels remain challenging to the learners by the end of training.

4.3.2 Auditory Discrimination

4.3.2.1 Tests

The model summary of the auditory discrimination (AD) task revealed statistically significant increases in the accuracy of responses during the mid-test ($p < 0.01$) and post-test ($p < 0.0001$) as compared to the pre-test, under the assumption that all other variables remained constant. The mid-test showed a probability of accurate answers of 70%, while the post-test showed a probability of accurate answers of 74%.

Table 4.11 presents a pairwise comparison analysis of tests (pre-, mid-, post-), aiming to thoroughly examine any potential disparities between them. An odds ratio of 0.832 between the mid- and post-tests indicates more improvement in the accuracy of responses during the post-test compared to the mid-test. Yet, the p-value exceeds the 0.05 threshold, which means that the variation in the correct responses between the mid- and post-tests cannot be conclusively claimed to be significant. On the other hand, the pre-test showed significantly lower accuracy in responses than both the mid-test, with an odds ratio of 0.430 ($p < 0.01$), and the post-test, with an odds ratio of 0.357 ($p < 0.01$). In other words, mid- and post-tests had a higher probability of accurate responses compared to the pre-test.

Contrast	***Odds.ratio***	***SE***	***z.ratio***	***p.value***
pre / mid	0.430	0.126	-2.880	0.0060
pre / post	0.357	0.102	-3.577	0.0010
mid / post	0.832	0.23	-0.661	0.5084

Table 4.11 Main effect of pre-, mid-, and post- tests for the AD task, excluding interactions. Odds ratio mid-point is 1.

The graphical representation below depicts the EMMeans of test accuracy of the AD task across the groups. The confidence intervals for the average predicted probabilities show that the pre-test has slightly more variability among the groups than the mid- and post-tests. The plot indicates that the performance of the groups consistently hit the ceiling effect across all test comparisons, as all rates are above 90%. Despite the initial ceiling effects, there was a modest increase in average predicted probability across the tests, with the post-test being slightly higher than the pre-test and mid-test.

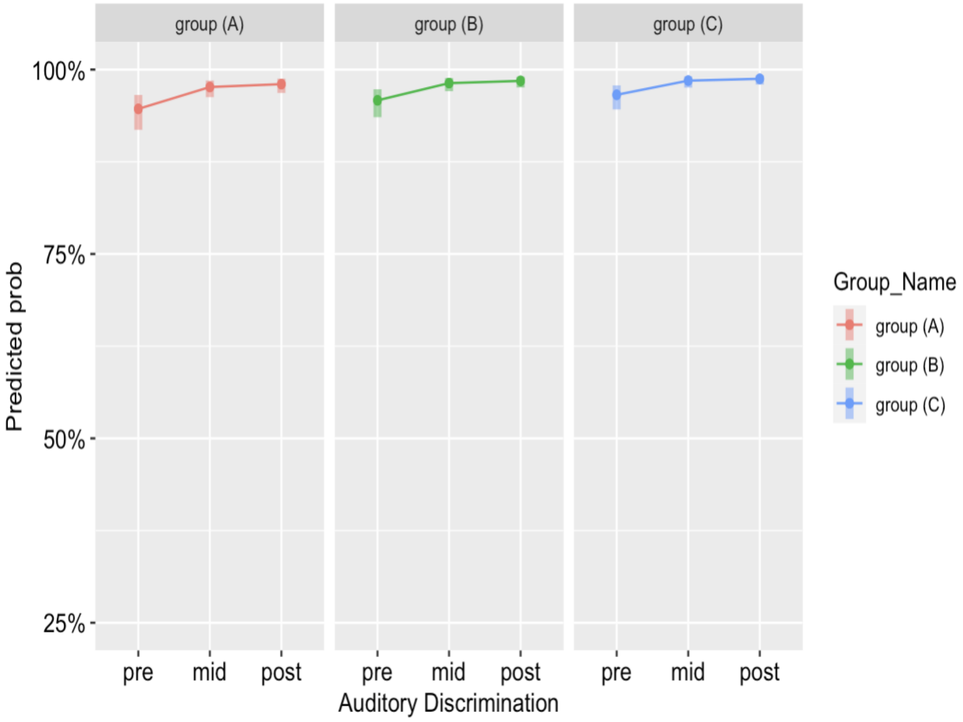


Figure 4.5 The EMMeans of pre-, mid-, and post-tests across groups for the AD task.

The interaction outcomes of the tests (pre, mid, and post) across the groups for the AD task are shown in Table 4.12 below[55]. Participants' performance in the post-test showed statistically significant improvement (3% to 4%, $p < 0.0001$) compared to the pre-test. Likewise, their performance on the mid-test was significantly better (by 2% to 3%, $p < 0.0001$) than on the pre-test. Marginal improvement was seen during

[55] To provide interactions, a multiple comparison t-test was conducted on predicted values derived from the raw data.

the post-test compared to the mid-test[56]. The post-test showed a marginal improvement over the mid-test by 1%. These outcomes indicate negligible differences across the tests, suggesting that the auditory discrimination task was straightforward for the learners.

Groups	*Contrast (test)*	*Mean Difference*	*p-value*
A	pre- mid	-3%	0.000
	pre-post	-4%	0.000
	mid- post	-1%	0.000
B	pre- mid	-3%	0.000
	pre-post	-3%	0.000
	mid- post	-0.4%	0.000
C	pre- mid	-2%	0.000
	pre-post	-3%	0.000
	mid- post	-1%	0.000

Table 4.12 [57] Interactions of pre-, mid-, and post- tests across groups for the AD task.

4.3.2.2 Groups

The results of the mixed-effect model analysis indicate that participants in group C showed a statistically significant improvement in response accuracy ($p < 0.05$) across the pre-, mid-, and post-tests, with a 61% higher chance of providing accurate responses compared to group A (the reference). On the other hand, the probability of group B delivering accurate responses did not exhibit statistical significance (56%, $p > 0.05$).

Table 4.13 sheds light on the results obtained from the pairwise comparison test of groups' performance. As shown, the accuracy of response did not show a significant

[56] Due to the use of raw data, the test comparisons (mid – post) revealed a high degree of statistical significance. Yet, the critical assessment lies in determining the extent of the mean difference.

[57] The mean difference is computed as the proportion of responses at the pre-test – proportion of responses at the mid-test or post-test. When the proportion at the mid-test (or post-test) is higher than that of the pre-test, this yields a negative mean difference. The same principle applies to *mid – post*

improvement between groups A and B (odds ratio = 0.773, *p* > 0.05) and between groups C and B (odds ratio = 0.813, *p* > 0.05). However, it was observed that group C displayed a tendency to achieve better performance compared to group A (odds ratio = 0.629, *p* = 0.0961).

Contrast	***Odds.ratio***	***SE***	***z.ratio***	***p.value***
group (A) / group (B)	0.773	0.161	-1.239	0.3228
group (A) / group (C)	0.629	0.136	-2.144	0.0961
group (B) / group (C)	0.813	0.175	-0.958	0.3380

Table 4.13 Main effect of groups (A, B, C) for the AD task across pre-, mid-, and post-tests, excluding interactions.

The following graph illustrates the EMMeans of group accuracy of the AD task across the test. The confidence intervals show that group A has more variability across the tests than groups C and B, with group B being more variable than group C. Across the three groups, a slight improvement is observed during the mid- and post-tests, attributable to the initial performance levels being near the ceiling, with all scores above 94%.

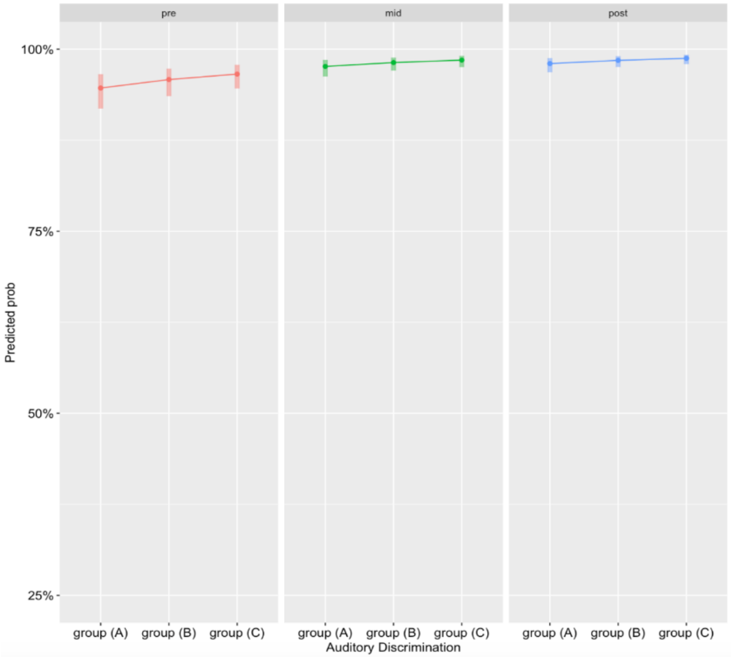


Figure 4.6 The EMMeans of groups (A, B, C) across the pre-, mid-, and post-tests for the AD task.

The interactions of groups across the tests (pre-, mid-, post-) for the AD task (see Table 4.14 below[58]) demonstrated that group B improved group A's performance by a negligible 1% across all three tests. Likewise, group C improved marginally compared to both groups A and B, by 0.39% to 2%. When considered as a whole, the performance of the three groups appears comparable, and the marginal improvement of groups B and C on the mid-and post-tests results from the marginal improvement observed on the pre-test.

Test	***Contrast (groups)***	***Mean Difference***	***p-value***
Pre-test	A – B	-1%	0.000
	A – C	-2%	0.000
	B – C	-1%	0.000
Mid-test	A – B	-1%	0.000
	A – C	-1%	0.000
	B – C	-0.39%	0.000
Post-test	A – B	-1%	0.000
	A – C	-1%	0.000
	B – C	-0.32%	0.000

Table 4.14 Interactions of groups across tests (pre-, mid-, post-) for the AD task.

4.3.2.3 Vowels

The multiple comparison t-test for the AD task revealed the *interactions* of tests (pre-, mid-, and post-) across vowels (See Table G.II in Appendix F). Compared to the pre-test, vowel performance significantly enhanced during the post-test (by 1% to 7%, $p <$ 0.0001) and the mid-test (by 1% to 6%, $p <$ 0.0001). However, there was only a slight increase of 0.1% to 0.9% in vowel performance from the mid-test to the post-test.

Table 4.15 shows the individual impact of vowels on response accuracy across the tests (*without interactions*). All vowels had a positive impact on response accuracy throughout the tests (scoring above the 50% threshold). While vowels were seemingly straightforward to discriminate before training during the pre-test, the process became even easier in the mid- and post-tests. The probability of accurate responses for the

[58] Due to the use of raw data, the group comparisons (A – B, A – C, B – C) revealed a high degree of statistical significance. Yet, the critical assessment lies in determining the extent of the mean difference.

vowels /e/ and /ɛə/ were 89% and 88%, respectively, in the pre-test; however, the accuracy increased by 6% during the mid-test and 7% during the post-test. The vowel /ɑː/accuracy was 93% at the pre-test and showed a 4% growth during the mid-and post-tests. In addition, the pre-test accuracy rate for the vowels /ɔː/, /ɒ/, /ɜː/, and /ʊ/ was 95%, and the accuracy rate for the vowel /e/ was 96%. The mid-and post-tests indicated a 3% improvement for these vowels. Similarly, the vowels /aɪ/, /aʊ/, /ʌ/, /ɪ/, and /əʊ/ had a high accuracy in the pre-test (97%), while increased by 2% in the mid- and post-tests. The vowels /ɔɪ/ and /iː/ also had a high probability of accurate responses on the pre-tests (98%), and there was a 1% improvement in the mid-and post-tests.

Vowels	***Pre-test***	***Mid-test***	***Post-test***
/ɔɪ/	97.9%	99.1%	99.2%
/iː/	97.6%	99.0%	99.0%
/aʊ/	97.4%	98.9%	99.1%
/əʊ/	97.4%	98.9%	99.0%
/ɪ/	97.0%	98.7%	98.9%
/ʌ/	96.8%	98.6%	98.8%
/aɪ/	96.7%	98.7%	98.9%
/æ/	95.9%	98.2%	98.5%
/uː/	95.7%	98.1%	98.4%
/eɪ/	95.6%	98.1%	98.4%
/ɔː/	95.3%	97.9%	98.3%
/ɒ/	95.2%	97.9%	98.2%
/ʊ/	94.6%	97.6%	98.0%
/ɜː/	94.8%	97.7%	98.1%
/ɑː/	93.5%	97.1%	97.6%
/e/	88.6%	94.8%	95.6%
/ɛə/	88.3%	94.6%	95.5%

Table 4.15 The EMMeans of vowels across pre-, mid-, and post-tests for the AD task[59].

[59] The correct probabilities are represented by the following colours: (dark green) for high, (medium green) for moderate, and (pale green) for low.

Figure 4.7 below visually illustrates the EMMeans of vowels across tests, along with their respective confidence intervals. While all vowels achieved scores above the 50% threshold, there is variation in terms of their confidence intervals. The vowels /ɛə/, /ɜː/, /ɑː/, /ɒ/, /e/, /eɪ/, and /ɔː/ stand out with greater confidence intervals, which point to increased fluctuations around their average predicted probabilities. The broader range in learners' perception of these vowels, especially in the pre-test, may indicate less certainty in their answers. In contrast, the remaining vowels (/æ/, /aɪ/, /aʊ/, /əʊ/, /iː/, /ɪ/, /ɔɪ/, /uː/, /ʊ/, /ʌ/) display narrower confidence intervals and demonstrate consistency in their perception, potentially indicating higher confidence levels.

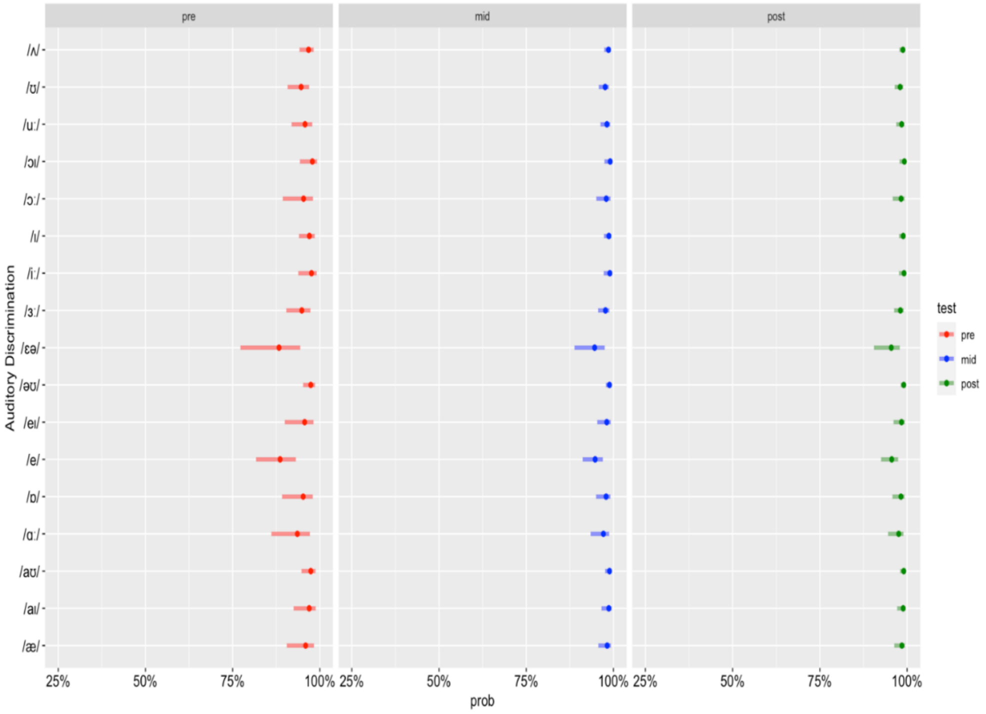


Figure 4.7 Probability of vowels across the pre-, mid-, and post-tests for the AD task.

In short, the findings suggest that within the auditory discrimination task context, each vowel was readily distinguishable during the pre-test, with all of them exceeding the 50% threshold. This underscores the fundamental simplicity of the task. Despite this, some progress was evident in both the mid-test and the post-test assessments. So, it can be deduced that the participants have successfully adjusted to the given task, thereby demonstrating the efficacy of the training.

4.3.2.4 Summary

The following points provide a summary of the results obtained from the AD task:

- Over time, there was an ongoing rise in performance seen across all groups, starting from the pre-test and advancing through the mid-test and post-tests for the AD task. This is consistent with the test's outcomes for the ID task.

- Groups B and C showed a modest increase in performance compared to group A, and group C marginally outperformed group B during the pre-test. The relatively increased performance seen in both the mid- and post-tests does not indicate better performance but simply reflects their initial performance. Consequently, all three groups demonstrated comparable levels of growth throughout the mid- and post-tests for the AD task. This aligns with the group's findings for the ID task.

- Incorporating training that includes variants of L1 accents alongside a single L2 variety, as well as training that features multiple L1 accents, proved non-detrimental for beginners. Comparative analysis revealed no substantive differences in performance between these groups and the group exposed to just one L1 variety when performing the AD task. These findings are consistent with the results observed in the ID task, where no substantive performance disparities were noted across the groups.

- During the ID task's pre-test, vowels classified as "easy" were marked by rates above 50% (/ɔɪ/, /eɪ/, /aʊ/, /æ/, /ɔː/, /ʌ/, /iː/, /aɪ/, and /ɑː/), whereas those deemed "difficult" had rates below 50% (/ɒ/, /e/, /ɪ/, /ɜː/, /ɛə/, /əʊ/, /ʊ/, and /uː/). Upon evaluating the AD task, it became evident that vowels were quite distinguishable, as they all received scores above 90% and reached the ceiling effect. Based on those findings, it can be determined that learners generally perceive the AD task as less demanding, thus considering it to be relatively effortless.

4.3.3 Category Discrimination

4.3.3.1 Tests

The analysis of the model's outcomes presented a noteworthy improvement in the likelihood of providing accurate responses. Specifically, during the mid-test, the accuracy rate reached 65% ($p < 0.01$), while in the post-test, it further increased to 73% ($p < 0.0001$) compared to the pre-test.

Table 4.16 clarifies the differences in accuracy observed across the tests, as described by pairwise test analysis. A significant difference was noted between the mid-test and post-test, wherein the post-test exhibited an increased level of accuracy compared to the mid-test (odd ratio= 0.665, $p < 0.05$). In a comparable manner, the mid and post-tests demonstrated significantly higher levels of accuracy compared to the pre-test (odd ratio= 0.547, $p < 0.01$), (odd ratio= 0.364, $p < 0.0001$).

contrast	***odds.ratio***	***SE***	***z.ratio***	***p.value***
pre / mid	0.547	0.1086	-3.038	0.0036
pre / post	0.364	0.0717	-5.130	< 0.0001
mid / post	0.665	0.1301	-2.086	0.0370

Table 4.16 The main effect of tests (pre-, mid-, post-) related to the CD task, excluding interactions. The mid-point of the odds ratio is 1.

The graphic shown below illustrates a gradual improvement in test performance across different groups for the CD task. The scores ranged from 34% to 42% during the initial assessment phase (pre-test). The mid-assessment scores ranged from 49% to 58%, indicating continuous improvement. This progress further increased during the post-test, where scores ranged from 59% to 67%. The stability of the confidence intervals for the average predicted probability of testing across the groups is evident.

The observed constancy in the confidence intervals suggests a sustained degree of precision in the estimation of the average test results.

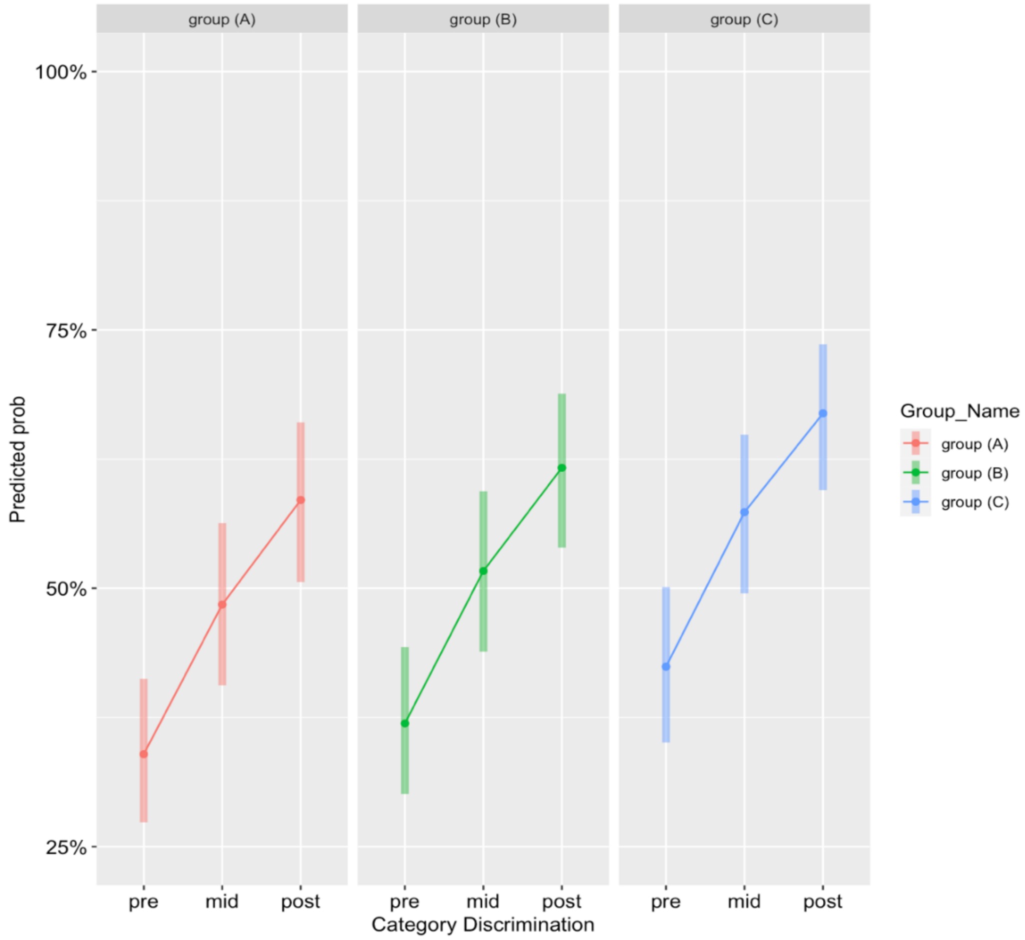


Figure 4.8 The EMMeans of pre-, mid-, and post-tests across the groups for the CD task (The threshold is 50%).

Interactions results of tests (pre-, mid-, and post-) across the groups for the CD task (see Table 4.17 below) show that groups A, B, and C performed significantly better on the mid-test (by 14%, p < 0.0001) and the post-test (by 23% to 24%, p < 0.0001) in comparison to the pre-test. In a similar vein, the mid-test results demonstrated a notable enhancement in comparison to the pre-test across all three groups (with an increase of 9% to 10%, p < 0.0001).

Groups	***Contrast (test)***	***Mean Difference***	***p-value***
A	pre- mid	-14%	0.000
	pre-post	-23%	0.000
	mid- post	-10%	0.000
B	pre- mid	-14%	0.000
	pre-post	-24%	0.000
	mid- post	-10%	0.000
C	pre- mid	-14%	0.000
	pre-post	-24%	0.000
	mid- post	-9%	0.000

Table 4.17 Interactions of tests (pre-, mid-, post-) across the groups for the CD task.

4.3.3.2 Groups

Using a mixed-effect model analysis, it was observed that group C demonstrated a significant level of response accuracy (p < 0.01), surpassing the performance of group A, with a correctness probability of 59% In comparison to group A, while group B's accuracy rate of 53% did not demonstrate statistical significance (p > 0.05).

In Table 4.18, a comprehensive examination of group performance is presented, employing pairwise comparison analysis. The interpretations of these results are discussed below:

- The odds ratio of 0.795 and the p-value of 0. 0712 indicate that group C tended to provide more accurate responses than group B.
- Although the odds ratio of 0.878 shows that group B outperformed group A, no statistically significant differences were found between the two groups (p > 0.05)
- The odds ratio of 0.69 and the p-value of 0.0073 (p < 0.01) indicate a statistically significant difference between groups A and C.

Contrast	*Odds.ratio*	*SE*	*z.ratio*	*p.value*
group (A) / group (B)	0.878	0.1024	-1.115	0.2648
group (A) / group (C)	0.698	0.0829	-3.030	0.0073
group (B) / group (C)	0.795	0.0921	-1.982	0.0712

Table 4.18 The main effect of the groups (A, B, C) in the CD task over the pre-, mid-, and post-tests, without considering interactions. The odds ratio's midpoint stands at 1.

Figure 4.9 displays the EMMeans that depict the probability of correct answers for all the groups across the pre-, mid-, and post-tests of the CD task. The consistency of the confidence intervals for the predicted average probability of groups A, B, and C within all assessments reflects a consistent degree of accuracy in the predictions of the average scores for each group. The figure illustrates a consistent pattern of improvement for all groups over time. The initial probabilities at the pre-test for groups A, B, and C were 34%, 37%, and 42% respectively. These climbed further in the mid- and post-tests.

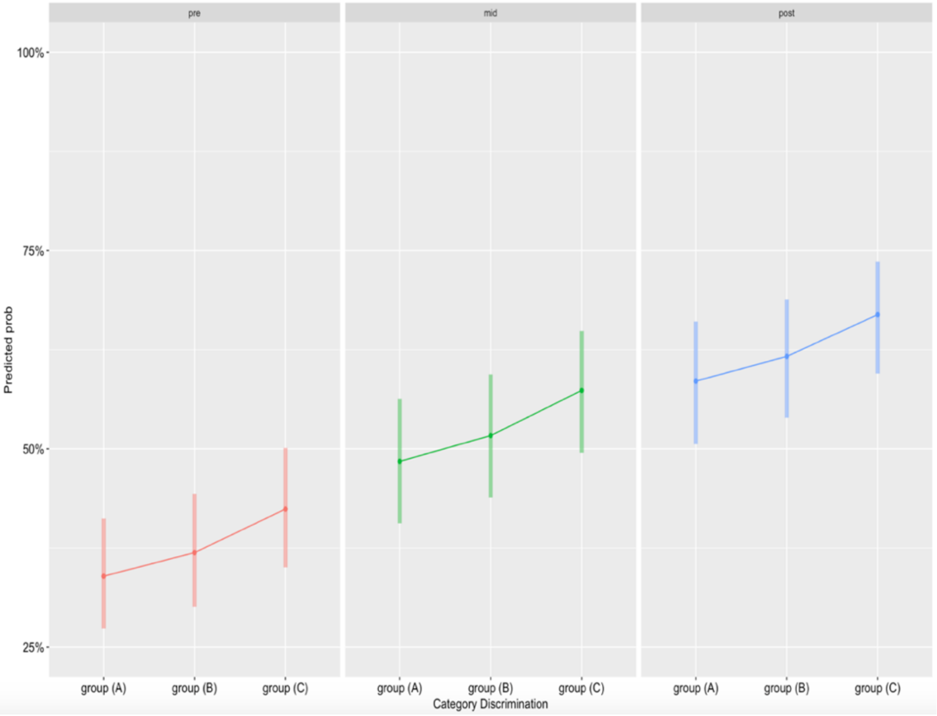


Figure 4.9 The EMMeans of groups (A, B, C) across the tests (pre-, mid-, post-) for the CD task (The threshold is 50%).

The results obtained from the multiple comparison t-test provide insights into the interactions among groups A, B, and C across the pre-, mid-, and post-test phases of the CD task (see Table 4.19). Based on the initial pre-test performance, it is evident that group C exhibited better performance compared to groups B and A (by 5% to 8%, *p* < 0.0001*).* However, the improved results observed for group C in the mid-test and post-test (5% to 9%) are not indicative of any additional progress. Rather, these findings reflect the participants' initial higher level of performance before the start of the training programme. Similarly, the higher score shown by group B, which had a 3% advantage over group A in both the mid- and post-tests, does not indicate a degree of progress. It merely reflects their pre-test performance, in which group B had a 3% edge over group A. Therefore, the performance of the three groups during the CD task is quite comparable.

Test	***Contrast (groups)***	***Mean Difference***	***p-value***
Pre-test	A - B	-3%	0.000
	A - C	-8%	0.000
	B - C	-5%	0.000
Mid-test	A - B	-3%	0.000
	A - C	-9%	0.000
	B - C	-6%	0.000
Post-test	A - B	-3%	0.000
	A - C	-8%	0.000
	B - C	-5%	0.000

Table 4.19 Interactions of groups across pre-, mid-, post-tests for the CD task.

4.3.3.3 Vowels

The multiple comparison t-test for the CD task demonstrated the interactions of tests (pre-, mid-, and post-) across vowels (See Table G.III in Appendix F). Compared to the pre-test, vowel performance showed a significant improvement during the post-test (by 16% to 25%, *p* < 0.0001) and the mid-test (by 12% to 16%, *p* < 0.0001). Furthermore, there was a notable enhancement in vowel performance during the mid-test compared to the pre-test, with an increase of 5% to 10% (*p* < 0.0001).

Table 4.20 presents the individual impacts of vowels on the probability of accurate responses during the pre-, mid- and post-test (without interactions). It appears there was difficulty distinguishing most vowels at the pre-test, as the accuracy rate was below the threshold of 50% probability. This can be attributed to the higher memory demands and greater uncertainty in the stimuli of the category discrimination task (Strange & Dittmann, 1984), which required learners to compare three distinct stimuli and discriminate the odd vowel among them. Specifically, during the pre-test, the vowels /ɜː/, /æ/, and /ɪ/ achieved the lowest probabilities, with scores of 21%, 22%, and 26% respectively. By the mid-test, the accuracy for /ɜː/ and /æ/ elevated by 12%, and for /ɪ/ by 13%. The accuracy of these vowels further improved during the post-test, with /ɜː/ and /æ/ both increasing by 21% and /ɪ/ by 23%. There was an 11% improvement in the post-test compared to the mid-test for these vowels.

Similarly, during the pre-test, the vowels /ʌ/, /aɪ/, /əʊ/, and /ɛə/ had a modest probability of 33% for correct responses, whereas the vowels /ɔː/ and /ɔɪ/ demonstrated a probability of 34%. In comparison to pre-test results, the accuracy of these vowels yielded a 14% improvement in the mid-test and further improved by 24% in the post-test. When comparing the results of the mid-test and post-test of these vowels, the latter showed a 10% improvement. Regarding the vowels /eɪ/, /uː/, and /ʊ/, each had a probability of 41% at the pre-test. Yet, the accuracy of these vowels increased by 15% in the mid-test, and this progress was further enhanced by a 25% improvement in the post-test. There was a 10% increase in vowel accuracy observed during the post-test in comparison to the performance demonstrated in the mid-test. Moreover, the accuracy rates for the vowels /ɒ/ and /iː/ in the pre-test were recorded as 43% and 47% respectively. The results of the mid-test indicated a 14% increase, whereas the post-test demonstrated a 24% improvement. In comparison to the mid-test, the post-test showed an additional 9% improvement in vowel accuracy.

In contrast, the vowel /ɑː/ stood out as the easiest to discriminate, providing a 70% probability of accurate answers in the pre-test. There was an 11% increase in the mid-test and a 16% rise in the post-test compared to the pre-test. A 5% improvement was also seen in the post-test scores compared to the mid-test scores. The vowel sounds /aʊ/ and /e/ exhibited a relatively high level of discriminability in the pre-test, with a 50% likelihood of correct responses. Progress was observed during the mid-test (15%) and post-test (23%) relative to the pre-test. As compared to the mid-test, the accuracy of vowels in the post-test was higher by 9%.

Interestingly, the pre-test results reveal notable differences in accuracy between the CD and ID tasks for the vowels /ɑː/, /e/, and /ɔɪ/. Specifically, the vowels /ɑː/ and /e/ had lower probabilities of correct answers in the ID task (51% and 36%, respectively) compared to the CD task (70% and 50%, respectively). This disparity may be attributed to the requirement in the ID task to identify /ɑː/ among various sets (clusters) of vowels (/ʌ/, /æ/, /ɑː/; /ɔː/, /ɑː/, /əʊ/; and /ɜː/, /ɑː/, /ɔː/) and /e/ among other sets of vowels (/iː/, /ɪ/, /e/; /e/, /eɪ/, /aɪ/; and /ɜː/, /ɛə/, /e/), potentially leading to orthographic confusion. In contrast, the CD task required distinguishing /ɑː/ only from /ɔː/ and /e/ only from /eɪ/, without the complication of spelling confusion, as participants did not see the words, resulting in higher accuracy. On the other hand, participants demonstrated a higher probability of correct responses for /ɔɪ/ in the ID task (69%) compared to the CD task (36%). This disparity might be due to the different cognitive demands of the tasks. In the ID task, participants only needed to identify /ɔɪ/ from among /aʊ/ and /əʊ/, leading to less orthographic confusion since no additional vowel sets were involved. In contrast, the CD task required participants to make finer distinctions between /aʊ/ and /ɔɪ/, necessitating the identification of the odd vowel /ɔɪ/ among three different words (foul, foul, foil). This higher cognitive load increased the difficulty, resulting in lower accuracy.

Vowels	***Pre-test***	***Mid-test***	***Post-test***
/ɑː/	70.4%	81.3%	86.7%
/aʊ/	50.0%	64.6%	73.3%
/e/	49.7%	64.3%	73.0%
/iː/	46.8%	61.7%	70.9%
/ɒ/	43.3%	58.2%	67.7%
/eɪ/	41.2%	56.1%	65.8%

/ʊ/	40.7%	55.7%	65.4%
/u:/	40.6%	55.5%	65.3%
/ɔ:/	34.3%	48.8%	58.9%
/ɔɪ/	33.5%	47.9%	58.1%
/ʌ/	32.8%	47.2%	57.3%
/əʊ/	33.2%	47.5%	57.7%
/aɪ/	32.6%	47.0%	57.1%
/ɛə/	31.6%	45.8%	56.0%
/ɪ/	26.1%	39.3%	49.3%
/æ/	21.5%	33.4%	43.0%
/ɜ:/	21.4%	33.3%	42.8%

Table 4.20 The EMMeans of vowels across pre-, mid-, and post-tests for the CD task[60]

The figure presented below graphically illustrates the EMMeans of vowels across tests during the CD task, excluding any interactions. The EMMeans are accompanied by their corresponding confidence intervals. The vowels /ɪ/, /ɛə/, /ɜ:/, /əʊ/, /ʌ/, /ʊ/, /u:/, /æ/, and /ɔ:/ demonstrated narrower confidence intervals, indicating lower degree of fluctuation around their average predicted probabilities. This suggests a greater level of certainty in the responses of learners. However, the vowels /ɑ:/, /aʊ/, /e/, /ɒ/, /i:/, /eɪ/, /ɔɪ/, and /aɪ/ display broader confidence intervals surrounding their mean predicted probabilities. This implies a comparatively lower level of confidence in the responses provided by learners. Recall that the vowels /ɑ:/, /aʊ/, and /e/ were most readily distinguished during the pre-test (/ɑ:/ scored 70%, while both /aʊ/ and /e/ scored 50%). Yet, given their wide confidence intervals, they seem to pose challenges for learners, suggesting a decrease in the certainty of their answers (greater individual variability).

[60] The correct probabilities are represented by the following colours: (dark green) for high, (medium green) for moderate, and (pale green) for low. Incorrect probabilities use (dark red) to indicate high, (medium red) to represent indicate, and (pale red) to indicate low.

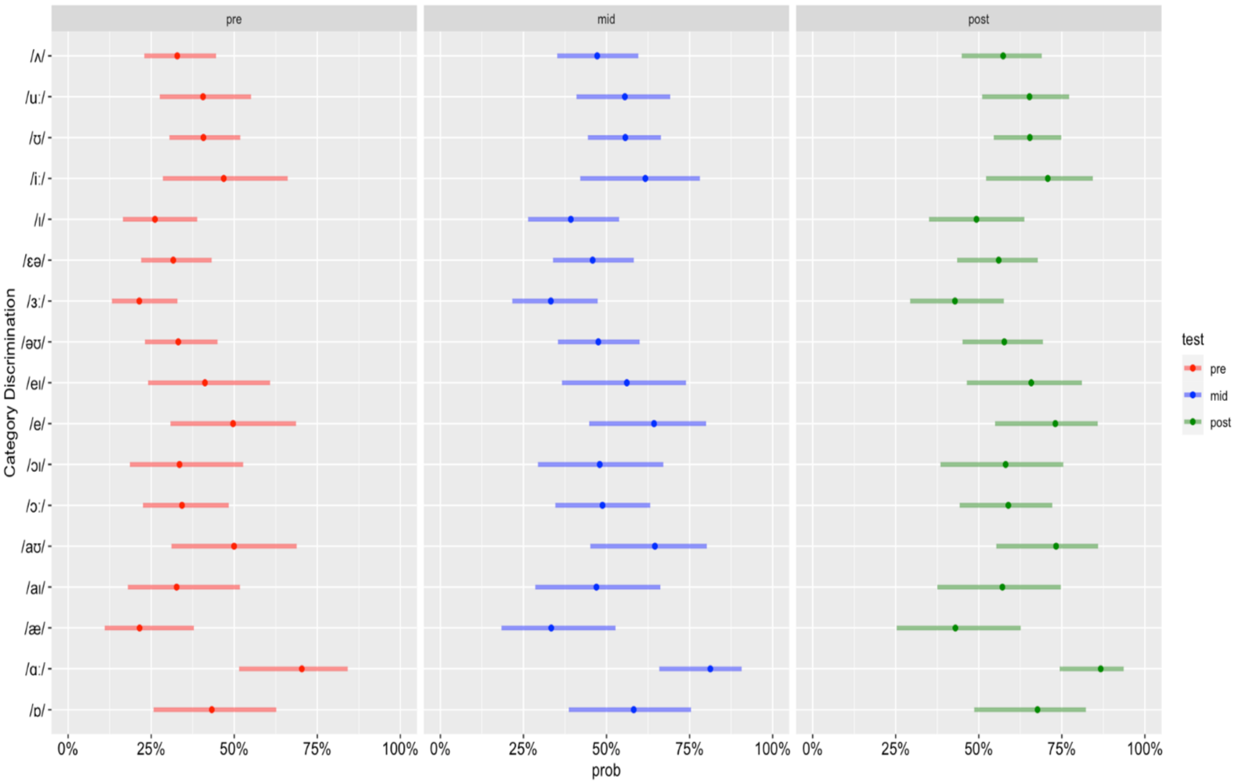


Figure 4.10 Probability of vowels across the pre-, mid-, and post-tests for the CD task.

In short, it appears that most vowels were difficult to discriminate at the pre-test while engaging in the category discrimination task. This greatly differs from the auditory discrimination task, which learners found fairly simple. Despite the inherent challenges of the category discrimination task, there was a discernible improvement during the mid-test, with even more significant progress observed in the post-test. This suggests that learners can adjust to the task at hand and perceive the training as effective, regardless of the difficulty associated with the task.

4.3.3.4 Summary

The subsequent points provide a summary of the outcomes of the CD task:

- During the CD task, there was an ongoing increase in performance across all groups, starting with the pre-test and continuing through the mid-test and post-tests. This corresponds to the test outcomes observed in both the ID and AD tasks.

- Group C does not exhibit greater improvement compared to groups A and B, and group B does not demonstrate greater improvement than group A during the mid-and post-tests. The performance of groups during the CD was quite comparable, which aligns with the group’s results seen in the ID and AD tasks.

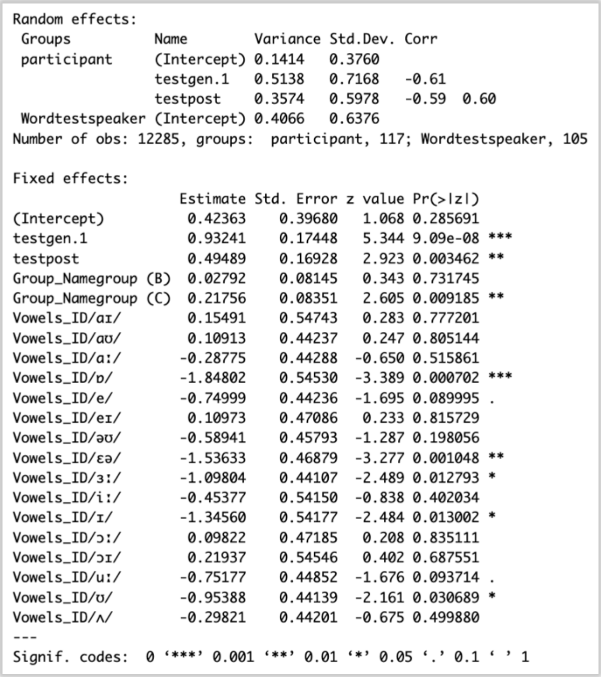
- The inclusion of training that incorporates various variations of the (L1) accents along with a single (L2) variety, as well as training that involves multiple L1 accents did not impede learning for beginners. The performance of these groups was comparable to that of the group exposed to only one L1 variety when engaging in the CD task. These observed results are consistent with the ID and AD task findings.

- During the pre-test phase of the CD task, learners perceived the large number of vowels to be particularly challenging. Only the vowels (/ɑː/, /aʊ/, and /e/) had a high probability of accurate responses (/ɑː/ scored 70%, whereas /aʊ/and /e/ both scored 50%). Despite this, the wide confidence intervals associated with these vowels indicate imprecision in the mean predicted probabilities, thereby increasing their perceived difficulty. The CD is significantly more difficult than the ID and AD tasks, which were rated as moderately difficult and relatively easy, respectively. Despite the difficulty of the CD task, vowel discrimination improved considerably at both the mid- and post-test stages, paralleling the upward trend observed in the ID and AD tasks. Therefore, the improvement in vowel accuracy across all three tasks indicates that the training is effective.

4.4 generalisation effects

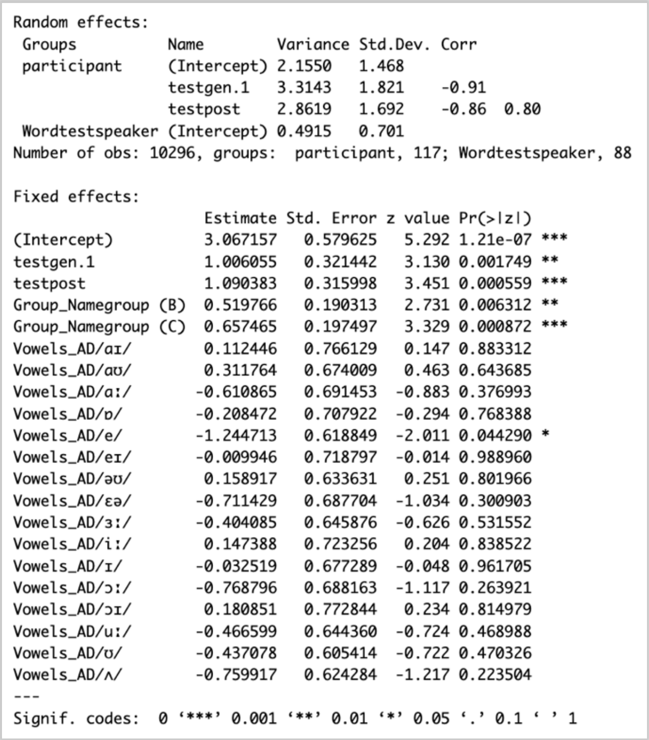
This section presents the results of two generalisation tests compared to pre- and post-tests for the ID, AD, and CD tasks. The first generalisation test (gen1) examines the effects of training on new items and speakers (1 SSBE, 1 Saudi) that were not previously encountered, while the second generalisation test (gen2) evaluates the effects of training on unfamiliar items (same stimuli as gen1 test) and speakers from both the inner and outer English circles (1 Indian, 1 Chinese). Tables 4.6 and 4.7 present the statistics for the comprehensive mixed-effects regression model, incorporating the primary fixed effects (test, group, and vowel) along with the random effects for the three perceptual tasks (ID, AD, CD). Table 4.21 focuses on the gen1 test in comparison to pre- and post-tests, while Table 4.22 focuses on the gen2 test in comparison to pre- and post-tests.

ID



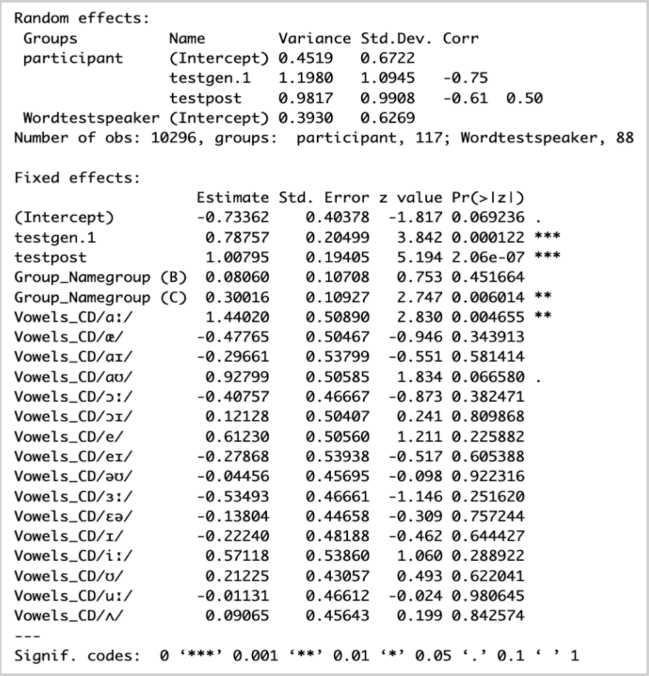
```
Random effects:
 Groups          Name        Variance Std.Dev. Corr
 participant     (Intercept) 0.1414   0.3760
                 testgen.1   0.5138   0.7168   -0.61
                 testpost    0.3574   0.5978   -0.59  0.60
 Wordtestspeaker (Intercept) 0.4066   0.6376
Number of obs: 12285, groups:  participant, 117; Wordtestspeaker, 105

Fixed effects:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           0.42363    0.39680   1.068 0.285691
testgen.1             0.93241    0.17448   5.344 9.09e-08 ***
testpost              0.49489    0.16928   2.923 0.003462 **
Group_Namegroup (B)   0.02792    0.08145   0.343 0.731745
Group_Namegroup (C)   0.21756    0.08351   2.605 0.009185 **
Vowels_ID/aɪ/         0.15491    0.54743   0.283 0.777201
Vowels_ID/aʊ/         0.10913    0.44237   0.247 0.805144
Vowels_ID/ɑː/        -0.28775    0.44288  -0.650 0.515861
Vowels_ID/ɒ/         -1.84802    0.54530  -3.389 0.000702 ***
Vowels_ID/e/         -0.74999    0.44236  -1.695 0.089995 .
Vowels_ID/eɪ/         0.10973    0.47086   0.233 0.815729
Vowels_ID/əʊ/        -0.58941    0.45793  -1.287 0.198056
Vowels_ID/ɛə/        -1.53633    0.46879  -3.277 0.001048 **
Vowels_ID/ɜː/        -1.09804    0.44107  -2.489 0.012793 *
Vowels_ID/iː/        -0.45377    0.54150  -0.838 0.402034
Vowels_ID/ɪ/         -1.34560    0.54177  -2.484 0.013002 *
Vowels_ID/ɔː/         0.09822    0.47185   0.208 0.835111
Vowels_ID/ɔɪ/         0.21937    0.54546   0.402 0.687551
Vowels_ID/uː/        -0.75177    0.44852  -1.676 0.093714 .
Vowels_ID/ʊ/         -0.95388    0.44139  -2.161 0.030689 *
Vowels_ID/ʌ/         -0.29821    0.44201  -0.675 0.499880
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

AD



```
Random effects:
 Groups           Name        Variance Std.Dev. Corr
 participant      (Intercept) 2.1550   1.468
                  testgen.1   3.3143   1.821    -0.91
                  testpost    2.8619   1.692    -0.86  0.80
 Wordtestspeaker (Intercept) 0.4915   0.701
Number of obs: 10296, groups:  participant, 117; Wordtestspeaker, 88

Fixed effects:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         3.067157   0.579625   5.292 1.21e-07 ***
testgen.1           1.006055   0.321442   3.130 0.001749 **
testpost            1.090383   0.315998   3.451 0.000559 ***
Group_Namegroup (B)  0.519766   0.190313   2.731 0.006312 **
Group_Namegroup (C)  0.657465   0.197497   3.329 0.000872 ***
Vowels_AD/aɪ/        0.112446   0.766129   0.147 0.883312
Vowels_AD/aʊ/        0.311764   0.674009   0.463 0.643685
Vowels_AD/ɑː/       -0.610865   0.691453  -0.883 0.376993
Vowels_AD/ɒ/        -0.208472   0.707922  -0.294 0.768388
Vowels_AD/e/        -1.244713   0.618849  -2.011 0.044290 *
Vowels_AD/eɪ/       -0.009946   0.718797  -0.014 0.988960
Vowels_AD/əʊ/        0.158917   0.633631   0.251 0.801966
Vowels_AD/ɛə/       -0.711429   0.687704  -1.034 0.300903
Vowels_AD/ɜː/       -0.404085   0.645876  -0.626 0.531552
Vowels_AD/iː/        0.147388   0.723256   0.204 0.838522
Vowels_AD/ɪ/        -0.032519   0.677289  -0.048 0.961705
Vowels_AD/ɔː/       -0.768796   0.688163  -1.117 0.263921
Vowels_AD/ɔɪ/        0.180851   0.772844   0.234 0.814979
Vowels_AD/uː/       -0.466599   0.644360  -0.724 0.468988
Vowels_AD/ʊ/        -0.437078   0.605414  -0.722 0.470326
Vowels_AD/ʌ/        -0.759917   0.624284  -1.217 0.223504
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

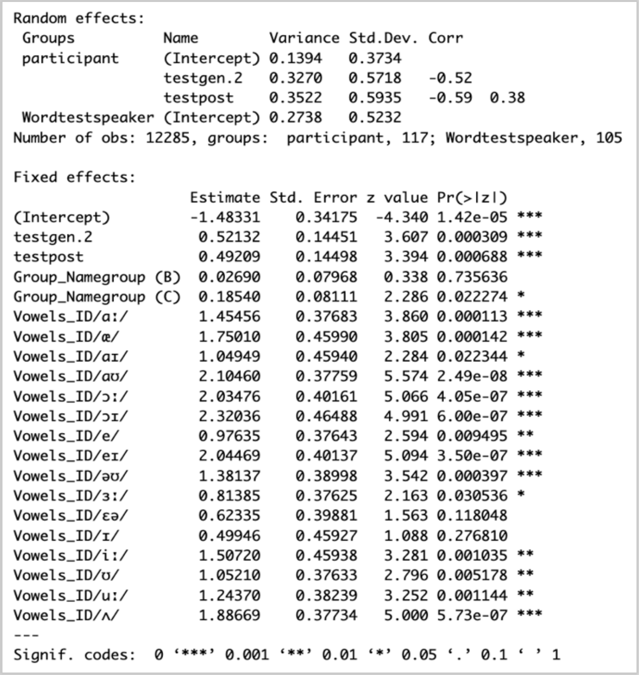
CD



```
Random effects:
 Groups           Name        Variance Std.Dev. Corr
 participant      (Intercept) 0.4519   0.6722
                  testgen.1   1.1980   1.0945   -0.75
                  testpost    0.9817   0.9908   -0.61  0.50
 Wordtestspeaker (Intercept) 0.3930   0.6269
Number of obs: 10296, groups:  participant, 117; Wordtestspeaker, 88

Fixed effects:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         -0.73362    0.40378  -1.817 0.069236 .
testgen.1            0.78757    0.20499   3.842 0.000122 ***
testpost             1.00795    0.19405   5.194 2.06e-07 ***
Group_Namegroup (B)  0.08060    0.10708   0.753 0.451664
Group_Namegroup (C)  0.30016    0.10927   2.747 0.006014 **
Vowels_CD/ɑː/        1.44020    0.50890   2.830 0.004655 **
Vowels_CD/æ/        -0.47765    0.50467  -0.946 0.343913
Vowels_CD/aɪ/       -0.29661    0.53799  -0.551 0.581414
Vowels_CD/aʊ/        0.92799    0.50585   1.834 0.066580 .
Vowels_CD/ɔː/       -0.40757    0.46667  -0.873 0.382471
Vowels_CD/ɔɪ/        0.12128    0.50407   0.241 0.809868
Vowels_CD/e/         0.61230    0.50560   1.211 0.225882
Vowels_CD/eɪ/       -0.27868    0.53938  -0.517 0.605388
Vowels_CD/əʊ/       -0.04456    0.45695  -0.098 0.922316
Vowels_CD/ɜː/       -0.53493    0.46661  -1.146 0.251620
Vowels_CD/ɛə/       -0.13804    0.44658  -0.309 0.757244
Vowels_CD/ɪ/        -0.22240    0.48188  -0.462 0.644427
Vowels_CD/iː/        0.57118    0.53860   1.060 0.288922
Vowels_CD/ʊ/         0.21225    0.43057   0.493 0.622041
Vowels_CD/uː/       -0.01131    0.46612  -0.024 0.980645
Vowels_CD/ʌ/         0.09065    0.45643   0.199 0.842574
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

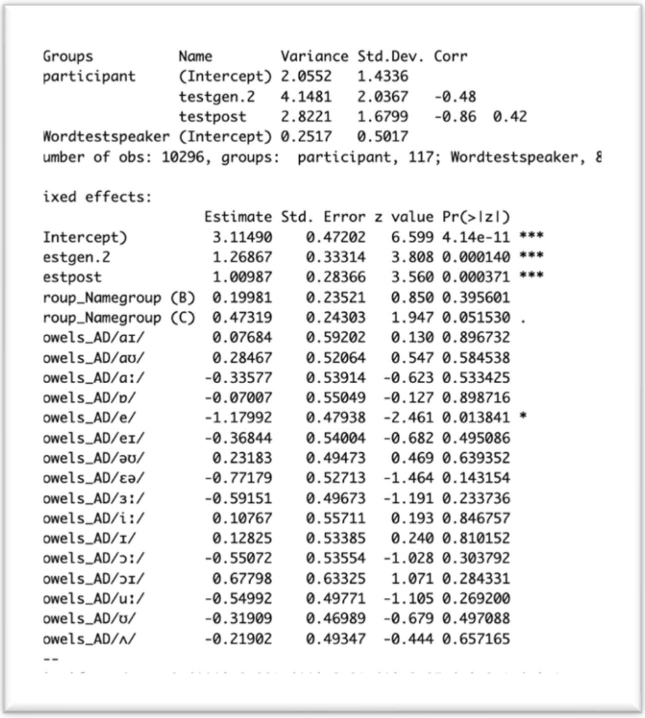
Table 4.21 Overview of model outputs, detailing fixed and random effects for ID, AD, and CD tasks (generalisation effects: gen1, pre-, and post- tests).

ID

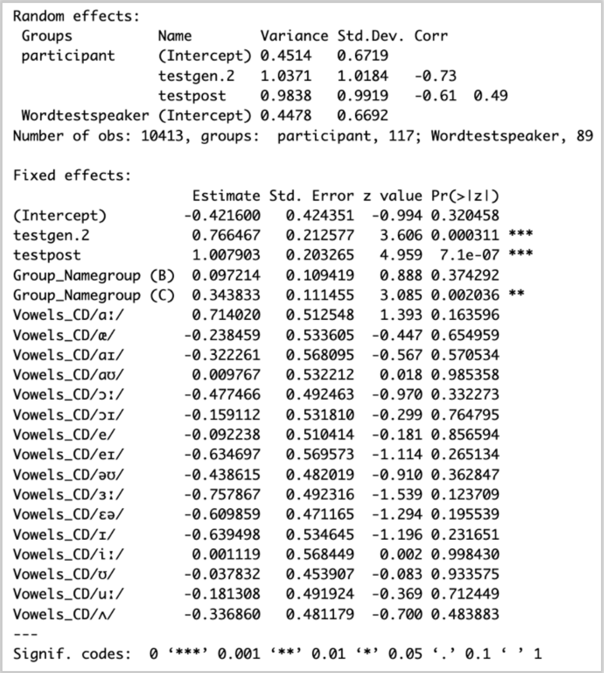


```
Random effects:
 Groups          Name        Variance Std.Dev. Corr
 participant     (Intercept) 0.1394   0.3734
                 testgen.2   0.3270   0.5718   -0.52
                 testpost    0.3522   0.5935   -0.59  0.38
 Wordtestspeaker (Intercept) 0.2738   0.5232
Number of obs: 12285, groups:  participant, 117; Wordtestspeaker, 105

Fixed effects:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         -1.48331    0.34175  -4.340 1.42e-05 ***
testgen.2            0.52132    0.14451   3.607 0.000309 ***
testpost             0.49209    0.14498   3.394 0.000688 ***
Group_Namegroup (B)  0.02690    0.07968   0.338 0.735636
Group_Namegroup (C)  0.18540    0.08111   2.286 0.022274 *
Vowels_ID/ɑː/        1.45456    0.37683   3.860 0.000113 ***
Vowels_ID/æ/         1.75010    0.45990   3.805 0.000142 ***
Vowels_ID/aɪ/        1.04949    0.45940   2.284 0.022344 *
Vowels_ID/aʊ/        2.10460    0.37759   5.574 2.49e-08 ***
Vowels_ID/ɔː/        2.03476    0.40161   5.066 4.05e-07 ***
Vowels_ID/ɔɪ/        2.32036    0.46488   4.991 6.00e-07 ***
Vowels_ID/e/         0.97635    0.37643   2.594 0.009495 **
Vowels_ID/eɪ/        2.04469    0.40137   5.094 3.50e-07 ***
Vowels_ID/əʊ/        1.38137    0.38998   3.542 0.000397 ***
Vowels_ID/ɜː/        0.81385    0.37625   2.163 0.030536 *
Vowels_ID/ɛə/        0.62335    0.39881   1.563 0.118048
Vowels_ID/ɪ/         0.49946    0.45927   1.088 0.276810
Vowels_ID/iː/        1.50720    0.45938   3.281 0.001035 **
Vowels_ID/ʊ/         1.05210    0.37633   2.796 0.005178 **
Vowels_ID/uː/        1.24370    0.38239   3.252 0.001144 **
Vowels_ID/ʌ/         1.88669    0.37734   5.000 5.73e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

AD



```
Groups          Name        Variance Std.Dev. Corr
participant     (Intercept) 2.0552   1.4336
                testgen.2   4.1481   2.0367   -0.48
                testpost    2.8221   1.6799   -0.86  0.42
Wordtestspeaker (Intercept) 0.2517   0.5017
umber of obs: 10296, groups:  participant, 117; Wordtestspeaker, 8

ixed effects:
                   Estimate Std. Error z value Pr(>|z|)
Intercept)          3.11490    0.47202   6.599 4.14e-11 ***
estgen.2            1.26867    0.33314   3.808 0.000140 ***
estpost             1.00987    0.28366   3.560 0.000371 ***
roup_Namegroup (B)  0.19981    0.23521   0.850 0.395601
roup_Namegroup (C)  0.47319    0.24303   1.947 0.051530 .
owels_AD/aɪ/        0.07684    0.59202   0.130 0.896732
owels_AD/aʊ/        0.28467    0.52064   0.547 0.584538
owels_AD/ɑː/       -0.33577    0.53914  -0.623 0.533425
owels_AD/ɒ/        -0.07007    0.55049  -0.127 0.898716
owels_AD/e/        -1.17992    0.47938  -2.461 0.013841 *
owels_AD/eɪ/       -0.36844    0.54004  -0.682 0.495086
owels_AD/əʊ/        0.23183    0.49473   0.469 0.639352
owels_AD/ɛə/       -0.77179    0.52713  -1.464 0.143154
owels_AD/ɜː/       -0.59151    0.49673  -1.191 0.233736
owels_AD/iː/        0.10767    0.55711   0.193 0.846757
owels_AD/ɪ/         0.12825    0.53385   0.240 0.810152
owels_AD/ɔː/       -0.55072    0.53554  -1.028 0.303792
owels_AD/ɔɪ/        0.67798    0.63325   1.071 0.284331
owels_AD/uː/       -0.54992    0.49771  -1.105 0.269200
owels_AD/ʊ/        -0.31909    0.46989  -0.679 0.497088
owels_AD/ʌ/        -0.21902    0.49347  -0.444 0.657165
--
```

CD	

```
Random effects:
 Groups          Name        Variance Std.Dev. Corr
 participant     (Intercept) 0.4514   0.6719
                 testgen.2   1.0371   1.0184   -0.73
                 testpost    0.9838   0.9919   -0.61  0.49
 Wordtestspeaker (Intercept) 0.4478   0.6692
Number of obs: 10413, groups:  participant, 117; Wordtestspeaker, 89

Fixed effects:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        -0.421600   0.424351  -0.994 0.320458
testgen.2           0.766467   0.212577   3.606 0.000311 ***
testpost            1.007903   0.203265   4.959  7.1e-07 ***
Group_Namegroup (B)  0.097214   0.109419   0.888 0.374292
Group_Namegroup (C)  0.343833   0.111455   3.085 0.002036 **
Vowels_CD/aː/       0.714020   0.512548   1.393 0.163596
Vowels_CD/æ/       -0.238459   0.533605  -0.447 0.654959
Vowels_CD/aɪ/      -0.322261   0.568095  -0.567 0.570534
Vowels_CD/aʊ/       0.009767   0.532212   0.018 0.985358
Vowels_CD/ɔː/      -0.477466   0.492463  -0.970 0.332273
Vowels_CD/ɔɪ/      -0.159112   0.531810  -0.299 0.764795
Vowels_CD/e/       -0.092238   0.510414  -0.181 0.856594
Vowels_CD/eɪ/      -0.634697   0.569573  -1.114 0.265134
Vowels_CD/əʊ/      -0.438615   0.482019  -0.910 0.362847
Vowels_CD/ɜː/      -0.757867   0.492316  -1.539 0.123709
Vowels_CD/ɛə/      -0.609859   0.471165  -1.294 0.195539
Vowels_CD/ɪ/       -0.639498   0.534645  -1.196 0.231651
Vowels_CD/iː/       0.001119   0.568449   0.002 0.998430
Vowels_CD/ʊ/       -0.037832   0.453907  -0.083 0.933575
Vowels_CD/uː/      -0.181308   0.491924  -0.369 0.712449
Vowels_CD/ʌ/       -0.336860   0.481179  -0.700 0.483883
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 4.22 Overview of model outputs, detailing fixed and random effects for ID, AD, and CD tasks (generalisation effects: gen2, pre-, and post- tests).

Section 4.4.1 below highlights the variability in how researchers report and interpret generalisation effects, justifying the approach of this study in comparing generalisation tests with pre- and post-tests. The subsequent subsections primarily examine the fixed effects (Tests, Groups, and Vowels) for the ID, AD, and CD tasks. The interpretation of results for all fixed factors is grounded in the concept of probability (refer to Section 4.2 for a detailed explanation). The approaches employed for analysing and reporting each fixed effect are as follows:

- A comprehensive explanation of the statistical summary for the mixed-effects regression model.
- Pairwise comparison tests (Wald tests) were conducted using the Estimated Marginal Means (EMMeans) package without interaction, accompanied by model-generated visualizations. Given that the raw data visualisations yielded similar conclusions, these figures are included in the appendix.
- Multiple comparison t-tests of predicted values based on the raw data with interaction.

4.4.1 Variability in reporting generalisation effects

The literature exhibits considerable variability in the methodologies researchers employ to report and interpret generalisation effects. Various approaches include:

- **Pre- to post-test comparisons:** Researchers such as Iverson et al. (2012) utilise the difference between pre- and post-test results to measure generalisation. This approach posits that improvements from pre- to post-test signify learning and the potential for generalisation.
- **Comparing generalisation tests with pre-test:** Others, like Brekelmans et al. (2022), compare the generalisation test with the pre-test. This method is employed to determine whether performance in generalisation tests significantly improves from pre-test levels, indicating successful generalisation.
- **Comparing generalisation tests with post-test:** Some studies, including Inceoglu (2014), compare generalisation tests with post-test results to assess the extent of generalisation.
- **Comparing generalisation tests with pre- and post-tests:** Researchers such as Carlet and Cebrian (2019) and Lengeris (2009) argue that generalisation is demonstrated if the generalisation results are as high as, or higher than, the post-test scores and significantly different from pre-test results.

Given this variability, the study considers the decision to compare generalisation tests with pre- and post-tests to provide a comprehensive view of distinct performance patterns:

- Stable pre- and post-test performance, followed by a decline in generalisation tests, indicates no retention or generalisation, as performance drops during generalisation tests.
- A gradual improvement from pre-test to post-test and then to generalisation tests signifies clear generalisation of skills.
- An initial increase during the pre-test, a decline in the post-test, and a subsequent rise in the generalisation test suggest that while post-test performance does not improve, it eventually aligns with pre-test levels.

- When post-test and generalisation test results are insignificantly different, but generalisation performance outperforms pre-test levels, it indicates significant improvements in generalisation compared to initial capabilities.

4.4.2 Vowel Identification

4.4.2.1 Tests

The model analysis showed a significant improvement in response accuracy in the gen1 ($p < 0.0001$) and post-test ($p < 0.01$) in comparison to the pre-test (the reference category). The gen1 test had a higher probability rate of 72% than the post-test's rate of 62%. In addition, the model analysis revealed a significant increase in response accuracy in the gen2 test and the post-test ($p < 0.0001$) compared to the pre-test. The gen2 test (63%) achieved a comparatively closer probability to the post-test (62%). These results suggest that the gen1 test had a greater probability than the post-test and the gen2 tests.

Table 4.23 shows a pairwise comparison analysis of the test, specifically highlighting the gen1 test in comparison to the pre and post-tests. It can be observed that gen1 demonstrated a significantly higher level of accuracy than both the pre-test (odds ratio = 0.394, $p < 0.0001$) and the post-test (odds ratio = 0.645, $p < 0.05$). Alternatively, Table 4.24 illustrates the comparative evaluation of gen2 with respect to the pre-and post-tests. The data demonstrate that gen2 achieved a significantly higher accuracy level than the pre-test (odds ratio of 0.594, $p < 0.001$), yet the accuracy remained almost the same when compared to the post-test (odds ratio of 0.970, $p > 0.05$).

Contrast	***Odds.ratio***	***SE***	***z.ratio***	***p.value***
pre / gen1	0.394	0.0687	-5.344	< .0001
post / gen1	0.645	0.2646	-2.561	0.0104

Table 4.23 Main effect of tests (pre-, post-, gen1) for the ID task, excluding interactions. Odds ratio mid-point is 1.

Contrast	***Odds.ratio***	***SE***	***z.ratio***	***p.value***
pre / gen.2	0.594	0.0858	-3.607	0.0009
post / gen2	0.970	0.1520	-0.198	0.8430

Table 4 24 Main effect of tests (pre-, post-, gen2) for the ID task, excluding interactions. Odds ratio mid-point is 1.

Figures 4.11 and 4.12 represent the EMMeans for the gen1 and gen2 tests of the ID task, respectively, in comparison to the pre-and post-tests within groups A, B, and C. The consistency of the CIs, which represents the average predicted probabilities for the tests in both figures, suggests a reliable degree of accuracy in anticipating the average test scores across all groups.

In Figure 4.11, a steady upward trend in the average predicted probability throughout the tests is apparent, with the gen1 test yielding higher values than both the pre-and post-tests. To be precise, the average probability of correct responses across the groups varied from 69% to 74% in the gen1 test. This is higher than the 47% to 52% seen in the pre-test and 59% to 64% seen in the post-test. These results imply that participants demonstrated proficiency in adapting to new items and speakers during the first generalisation test for the ID task.

On the other hand, Figure 4.12 shows that the average probability of correct responses across the groups for the gen2 test, ranging from 59% to 64%, was higher than the pre-test range of 46% to 50%, yet relatively similar to the values from the post-test, which fluctuated from 59% to 63%. The consistent outcomes in both the post-test and gen2 demonstrate that participants effectively acquired and applied the skills they learned during the training. This enhanced perception likely assisted them in unfamiliar words and speakers.

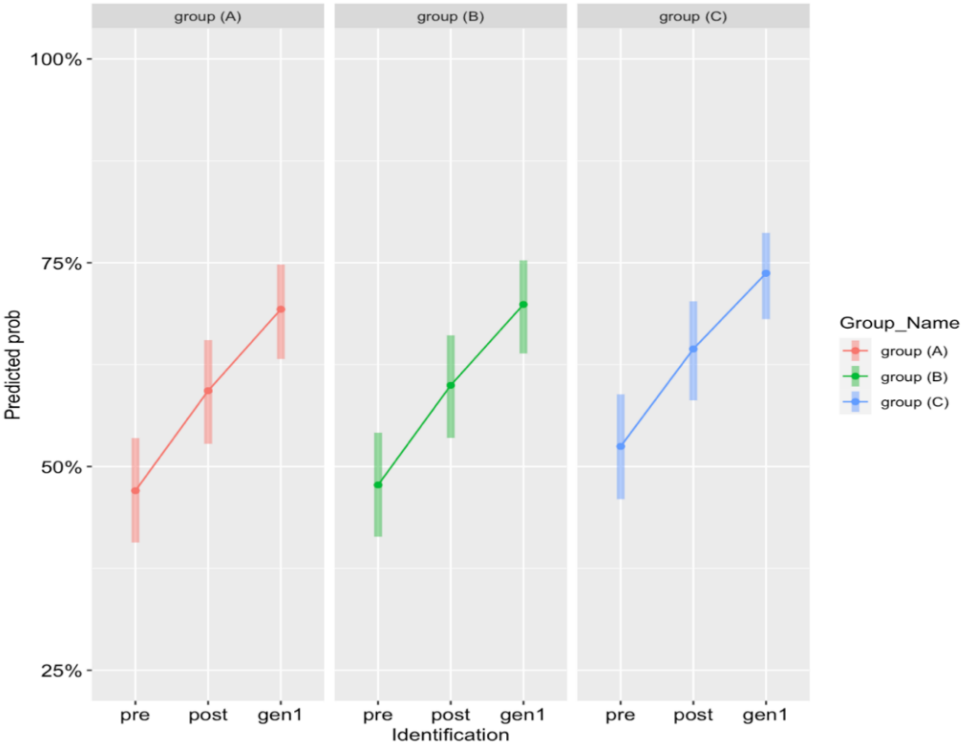


Figure 4.11 The EMMeans of gen1 compared to pre- and post-tests across groups for the ID task.

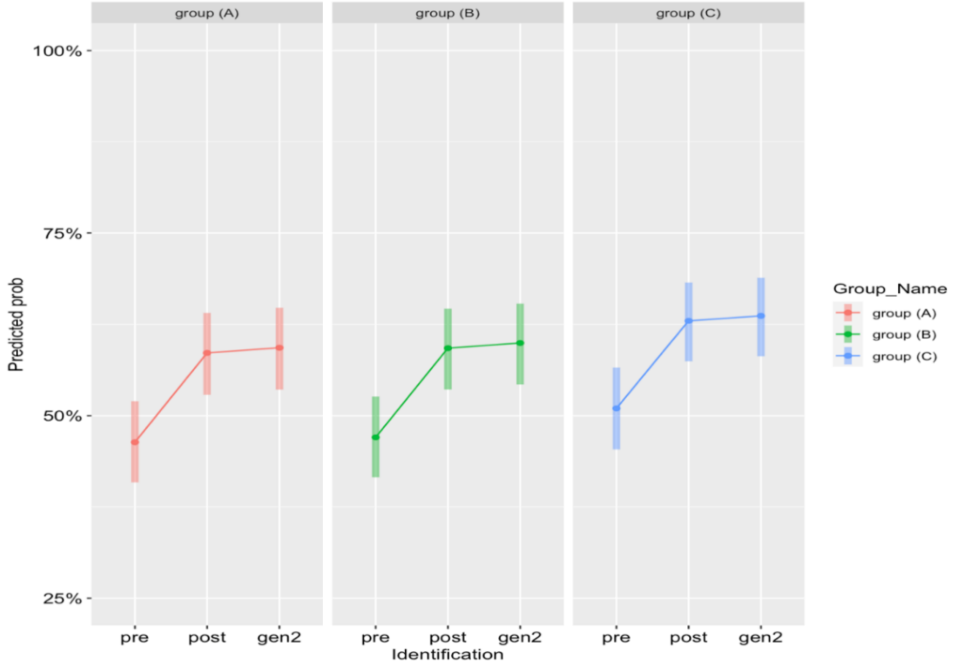


Figure 4.12 The EMMeans of gen2 compared to pre- and post-tests across groups for ID task.

The interactions of tests (gen1 and gen2 in relation to pre-and post-tests) across the groups were determined using the multiple comparison t-test (see Tables 25 and 26 below). The results revealed that groups' performance on the gent 1 was statistically higher than both the pre-test (by 20% to 21%, *p* < 0.0001) and the post-test (by 9% to 10%, *p* < 0.0001). As well, the performance of all three groups on gen2 was considerably better than their performance on the pre-test (by 21%, *p* < 0.0001). However, there was no difference between their performance on the gen2 test and the post-test.

Groups	***Contrast***	***Mean Difference***	***p-value***
A	pre - post	-11%	0.000
	pre - gen.1	-21%	0.000
	post- gen.1	-10%	0.000
B	pre - post	-11%	0.000
	pre - gen.1	-21%	0.000
	post- gen.1	-10%	0.000
C	pre - post	-11%	0.000
	pre - gen.1	-20%	0.000
	post- gen.1	-9%	0.000

Table 4.25 Interactions of tests (gen1, pre-, post-) across the groups for the ID task.

Groups	***Contrast***	***Mean Difference***	***p-value***
A	pre - post	-11%	.000
	pre - gen.2	-12%	.000
	post- gen.2	-0.75%	0.13
B	pre - post	-11%	.000
	pre - gen.2	-12%	.000
	post- gen.2	-0.75%	0.12
C	pre - post	-11%	.000
	pre - gen.2	-12%	.000
	post- gen.2	-0.72%	0.14

Table 4.26 Interactions of tests (gen2, pre-, post-) across the groups for the ID task.

4.4.2.2 Groups

The coefficients derived from the mixed-effect models provided insight into the impact of the gen1 and gen2 tests on group performance for the ID task, considering both pre-test (the test's reference category) and post-test outcomes. With the inclusion of the gen1, pre-and post-tests, there was a statistically significant positive effect on the accuracy of responses for group C (55%, $p < 0.01$), while the probability of correct responses within group B did not demonstrate a significant effect (51%, p 0.05) compared to group A (the reference category of groups). Similarly, when analysing the results of the model with the inclusion of the gen2, pre-, and post-tests, group C participants continued to demonstrate notably improved performance (p <0.05) with a probability of 55%, whereas group B participants showed a probability of 51% ($p >$ 0.05) compared to group A.

Table 4.27 provides a pairwise comparison test that evaluates the performance of each group in regard to the model comprising gen1, pre-and post-tests. There was a noticeable statistical disparity between groups A and C (odds ratio= 0.804, $p < 0.05$) and between groups B and C (odds ratio= 0.827, $p < 0.05$), where group C scored higher than both groups A and B. However, no significant difference was found between groups A and B (odds ratio= 0.972, $p > 0.05$).

Contrast	***Odds.ratio***	***SE***	***z.ratio***	***p.value***
group (A) / group (B)	0.972	0.0792	-0.343	0.7317
group (A) / group (C)	0.804	0.0672	-2.605	0.0276
group (B) / group (C)	0.827	0.0686	-2.287	0.0333

Table 4.27 Main effect of groups across the pre-, post-, and gen1-tests for the ID task.

In contrast, Table 4.28 displays a pairwise comparison test that assesses the performance of each group in relation to the model that includes gen2, pre-and post-tests. Group C showed a tendency for a greater probability of providing accurate answers compared to both group A (odds ratio = 0.831, *p* = 0.0668) and group B (odds ratio= 0.853, p = 0.0717). However, there was no significant discrepancy in performance between groups A and B (odds ratio= 0.973, *p* > 0.05).

Contrast	*Odds.ratio*	*SE*	*z.ratio*	*p.value*
group (A) / group (B)	0.973	0.0776	-0.338	0.7356
group (A) / group (C)	0.831	0.0674	-2.286	0.0668
group (B) / group (C)	0.853	0.0683	-1.979	0.0717

Table 4.28 Main effect of groups across the pre-, post-, and gen2-tests for the ID task.

Figure 4.13 and Figure 4.14 display the EMMeans, illustrating the probabilities of correct responses by the groups during the gen1 and gen2 tests, respectively, considering the pre-and post-test results. The CIs for the predicted average probabilities of groups A, B, and C throughout all evaluations demonstrate steadiness, suggesting a consistent level of accuracy in predicting each group's average scores.

In Figure 4.13, group C demonstrated a somewhat higher likelihood of correct responses during the gen1 test (74%) compared to groups A (69%) and B (70%). Group C also demonstrated better performance in the pre-test (52%) and post-test (64%) compared to group A (47%, 59%) and group B (47%, 60%). Similarly, during the gen2 test (see Figure 4.14), group C had a slightly higher likelihood of correct responses (64%) than group A (59%) and B (60%). This pattern persisted in both the pre-test and post-test, with group C (51%, 63%) outperforming the performance of groups A (46%, 59%) and B (47%, 59%). However, given that group C exceeds groups A and B in the pre-test before training, it can be inferred that all three groups demonstrate similar performance in the two generalisation tests.

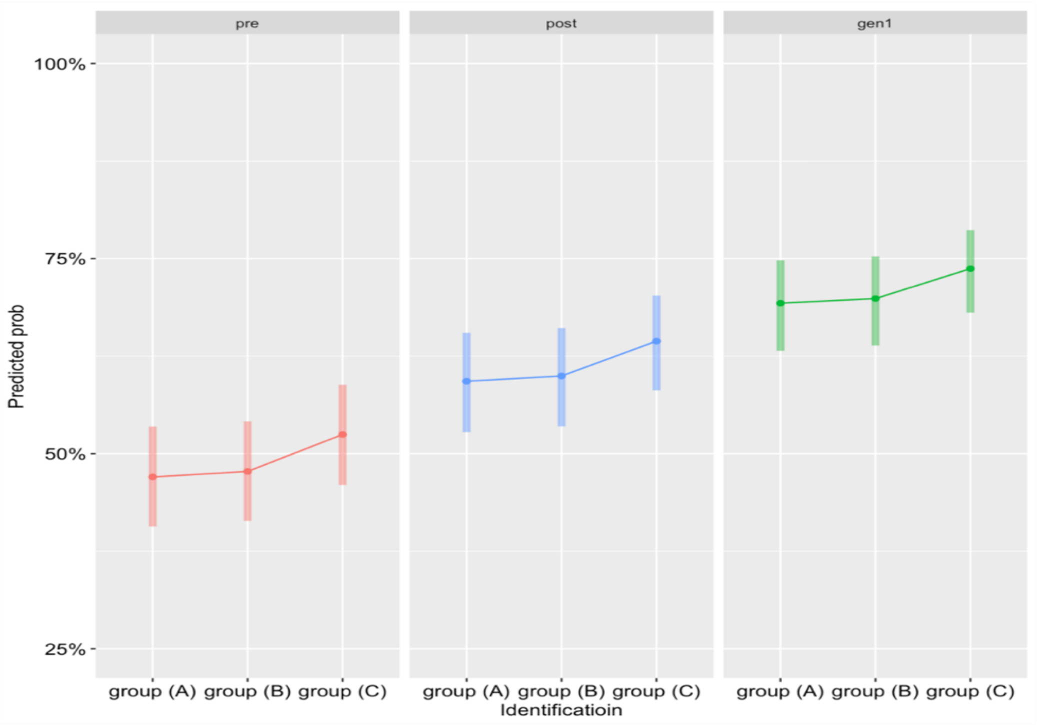


Figure 4.13 The EMMeans of groups (A, B, C) considering the gen1 in relation to pre-and post-test for the ID task.

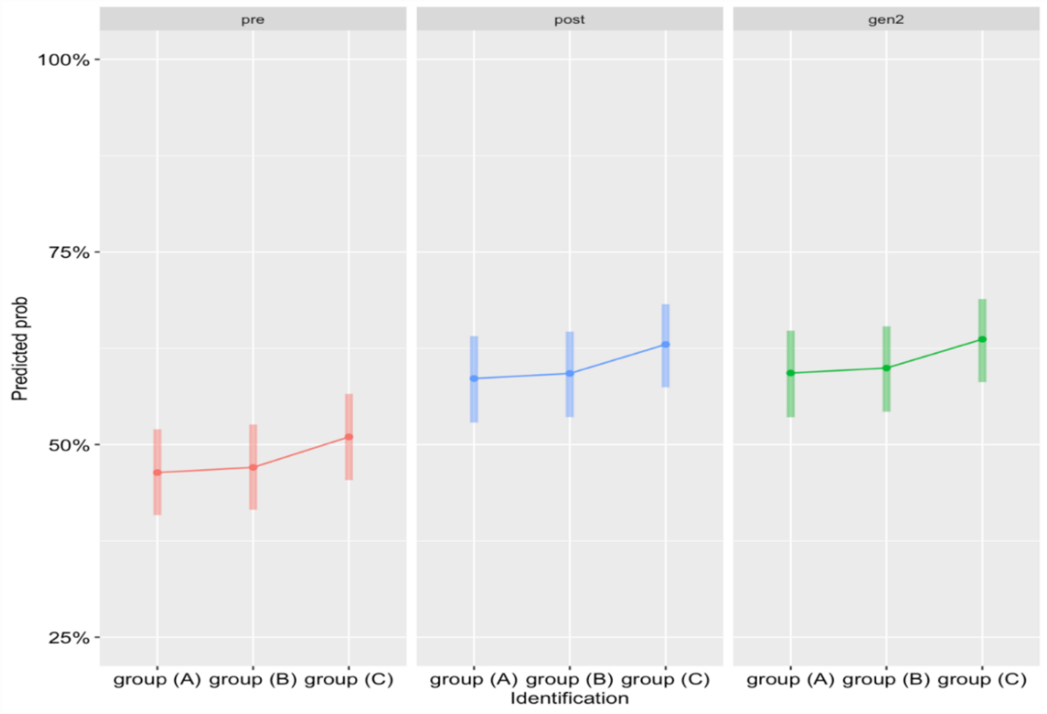


Figure 4.14 The EMMeans of groups (A, B, C) considering the gen2 in relation to pre-and post-test for the ID task.

The multiple comparison t-test was performed to determine the **interactions** of groups across the tests (gen1, gen2) in relation to the pre-and post-test for the ID task (see the tables below[61]). There were no notable differences between groups A and B across all assessments. Group C, on the other hand, showed considerably better performance compared to groups A and B during gen1 gen2 tests (by 4% to 5%, $p < 0.0001$). This observed improvement in the performance of participants in group C during the generalisation tests does not directly indicate a greater level of performance in comparison to other groups. This is due to the significant performance improvement shown by this group during the pre-test, exceeding the results of groups A and B by 4% to 5% ($p < 0.0001$).

Test	*Contrast (groups)*	*Mean Difference*	*p-value*
Gen1	A - B	-1%	0.2
	A - C	-4%	0.000
	B - C	-4%	0.000
Pre-test	A - B	-1%	0.19
	A - C	-5%	0.000
	B - C	-4%	0.000
Post-test	A - B	-1%	0.2
	A - C	-5%	0.000
	B - C	-4%	0.000

Table 4.29 Interactions of groups across tests (gen1, pre-, post-) for the ID task.

Test	*Contrast (groups)*	*Mean Difference*	*p-value*
Gen2	A - B	-1%	0.21
	A - C	-5%	0.000
	B - C	-4%	0.000
Pre-test	A - B	1%	0.21
	A - C	-5%	0.000
	B - C	-4%	0.000
Post-test	A - B	-1%	0.21
	A - C	-4%	0.000
	B - C	-4%	0.000

Table 4.30 Interactions of groups across tests (gen2, pre-, post-) for the ID task.

[61] The mean difference is derived by subtracting the response proportions of group B from group A and subtracting group C's proportion from both A and B across the tests. With group C's proportion exceeding that of groups A and B, and group B's proportion surpassing A's, the mean difference appears negative.

4.4.2.3 Vowels

The multiple comparison t-test was performed to determine the interactions of tests (gen1, gen2) in relation to the pre-and post-test across the vowels for the ID task (Refer to Tables G.IV and G.V in Appendix F). A significant improvement in vowel accuracy was observed during gen1 (17% to 23%, p < 0.0001) and gen2 (9% to 13%, *p* < 0.0001) when compared to the pre-test. Similarly, a notable improvement in vowel performance was seen during gen1 (9% to 13%, *p* < 0.0001) and gen2 (4% to 7%, *p* < 0.0001), relative to the post-test.

Table 4.31 and Table 4.32 showcase the individual impact of gen1 and gen2 tests on the probability of correct responses in comparison to the pre-and post-tests (without interactions). Table 4.31 reveals that while the vowels /æ/, /aɪ/, /aʊ/, /ɑː/, /eɪ/, /iː/, /ɔː/, /ɔɪ/, and /ʌ/ had an accuracy rate of over 50% in both the pre-and post-tests, the gen1 test showed an even higher likelihood of accuracy. Specifically, the accuracy of responses for the vowels /aɪ/, /aʊ/, /ɔː/, and /ɔɪ/ improved by 17% in the gen1 test compared to the pre-test and saw a 7% rise when compared to the post-test. The response accuracy for the vowels /ɑː/ and /ʌ/ increased by 21% in the gen1 test relative to the pre-test, and it was 9% higher when compared to the post-test. Moreover, the gen1 test showed an 18% increase in accurate answer rates for /eɪ/ and /æ/ and a 22% increase for /iː/ compared to the pre-test. These vowels also exhibited enhancements in their response accuracy by 7%, 8%, and 10%, respectively, when compared to the post-test.

However, the vowels /ɒ/, /ɛə/, /ɜː/, /ɪ/, /e/, /əʊ/, /uː/, and /ʊ/ exhibited a probability of correct responses below 50% in the pre-test. However, it is interesting to note that vowel accuracy during the gen1 test revealed a notable improvement compared to pre-tests as well as the post-tests. In particular, the accuracy of responses for the /ɒ/ and /ɛə/ vowels increased by 19% and 22%, respectively, when compared to the results obtained from the pre-test. Further, there was a 10% increase observed for the /ɒ/ and an 11% increase for the /ɛə/ when comparing the post-test results. When considering the vowels /uː/, /ʊ/ and /e/, there was a 23% improvement in accuracy in the gen1 test compared to the pre-test and a commendable improvement of 11% in comparison to the post-test. With respect to the vowels /ɜː/, /ɪ/, and /əʊ/, a significant improvement of 22% was observed during the gen1 test compared to the pre-test. Furthermore, a notable 10% improvement was observed when comparing the results to the post-test. In summary, the results reveal that participants were consistently able to retain the knowledge gained during the training for the identification task on the gen1 test. The accuracy of vowels during the gen1 test surpassed both the pre-test and post-test. The progression of vowels was approximately twice as noticeable when the gen1 test was contrasted with the pre-test rather than the post-test. This improvement was evident across all vowel types, including the simpler vowels (/æ/, /aɪ/, /aʊ/, /ɑː/, /eɪ/, /iː/, /ɔː/, /ɔɪ/, and /ʌ/) as well as the more challenging ones (ɒ/, /ɛə/, /ɜː/, /ɪ/, /e/, /əʊ/, /uː/, /ʊ/).

Vowels	***Pre-test***	***Post-test***	***Gen1***
/ɔɪ/	67.4%	77.2%	84.0%
/aɪ/	65.9%	76%	83.1%
/aʊ/	64.9%	75.2%	82.4%
/ɔː/	64.6%	75.0%	82.3%
/eɪ/	64.9%	75.2%	82.5%
/æ/	62.4%	73.1%	80.8%
/ʌ/	55.2%	66.9%	75.8%
/ɑː/	55.4%	67.1%	76.0%
/iː/	51.3%	63.3%	72.8%
/əʊ/	47.9%	60.1%	70.0%
/uː/	43.9%	56.2%	66.5%
/e/	43.9%	56.2%	66.5%
/ʊ/	39.0%	51.2%	61.9%
/ɜː/	35.6%	47.6%	58.4%
/ɪ/	30.2%	41.5%	52.3%
/ɛə/	26.3%	36.9%	47.5%
/ɒ/	20.7%	30.0%	39.9%

Table 4.31 The EMMeans of vowels across tests (gen1 compared to pre- and post-test) for the ID task[62]

[62] The correct response probabilities are represented by the following colours: (dark green) for high, (medium green) for moderate, and (pale green) for low. Incorrect response probabilities use (dark red) to indicate high, (medium red) to represent indicate, and (pale red) to indicate low.

On the other hand, Table 4.32 demonstrates that the vowels /ɑː/, /æ/, /aʊ/, /ɔː/, /ɔɪ/, /eɪ/, /iː/, and /ʌ/ achieved a high degree of accuracy (with a probability of greater than 50% for correct answers) not only in the pre-and post-tests but also in the gen2 test. A higher improvement in vowel accuracy was observed during the gen2 compared to the pre-test. Specifically, there was a 10% increase observed for /ɔɪ/, an 11% increase for /ʌ/ and /ɔː/and a 13% improvement for /iː/ and /ɑː/. Compared to the post-test outcomes, it was seen that these vowels demonstrated a slight improvement of 1% in the gen2 test. Similarly, the vowels /aʊ/, /eɪ/, and /æ/ displayed a higher accuracy rate in the gen2 test compared to the pre-test, showing increases of 10%, 11%, and 12% respectively. The accuracy of these vowels in the post-test was indistinguishable from that in the gen2 test (less than 1%).

Although all three assessments (pre-, post-, and gen2) showed below 50% accuracy for the challenging vowel /ɒ/, gen2 led a 9% gain in accuracy compared to pre-test, while the differences between gen2 and post-test were the same. Likewise, the vowels /ɜː/, /ɛə/, and /ɪ/ had low accuracy (below 50%) across the three tests; however, an advancement was observed for the vowels /ɜː/, /ɛə/ (12%) and the vowel /ɪ/ (11%) in the gen2 test when compared to the pre-test, while their accuracy relative to the post-test stayed the same (difference is just 1%). The accuracy improvement of the vowels /aɪ/, /e/, /əʊ/, /ʊ/, and /uː/ grew by 13% in the gen2 test relative to the pre-test, while the progress was almost identical to the post-test, with a mere 1% increase.

To conclude, the accuracy of all vowels for the identification task on the gen2 test, both those perceived as easy and difficult, was better than the performance on the pre-test, while it mirrored the post-test results. These findings further indicate an apparent retention of knowledge, as the identification of vowels during the gen2 test showed no decline compared to the post-test. Examining the performance of vowels in the ID task throughout gen1 and gen2, the data clearly demonstrate that the gen2 test is indicative of retention, whereas gen1 shows ongoing improvement.

Vowels (ID)	***Pre-test***	***Post-test***	***Gen 2***
/ɔɪ/	71.3%	80.2%	80.7%
/aʊ/	66.7%	76.6%	77.1%
/eɪ/	65.3%	75.5%	76.0%
/ɔː/	65.1%	75.3%	75.8%
/ʌ/	61.6%	72.4%	73.0%
/æ/	58.4%	69.6%	70.2%
/iː/	52.4%	64.3%	64.9%
/ɑː/	51.1%	63.0%	63.7%
/əʊ/	49.2%	61.3%	62.0%
/uː/	45.8%	58.0%	58.7%
/ʊ/	41.1%	53.3%	54.0%
/aɪ/	41.0%	53.2%	53.9%
/e/	39.3%	51.4%	52.1%
/ɜː/	35.5%	47.3%	48.1%
/ɛə/	31.2%	42.6%	43.3%
/ɪ/	28.6%	39.6%	40.3%
/ɒ/	19.6%	28.5%	29.1%

Table 4.32 The EMMeans of vowels across tests (gen2 compared to pre- and post-test) for the ID task.

Figures 4.15 and 4.16 present a visual representation of the EMMeans for vowel responses observed during the gen1 and gen2 tests, considering the results of both pre-and post-tests. As demonstrated, the average probability of responding correctly in the gen1, gen2, and post-tests was higher compared to that observed in the pre-test. Vowel accuracy showed improvement (more probability of accurate answers) in gen1 in contrast to the post-test while maintaining similar levels during gen2 when compared to the post-test. The vowels /iː/, /ɪ/, /aɪ/, and /æ/ exhibit broader CIs, implying a greater fluctuation around their average predicted probabilities. This could be interpreted as a reflection of individual differences in learners' perception of these vowels. In contrast, the rest of the vowels demonstrate smaller CIs, showing lesser fluctuation around their average predicted probabilities. This might hint that learners perceive these vowels more uniformly, indicating higher confidence levels.

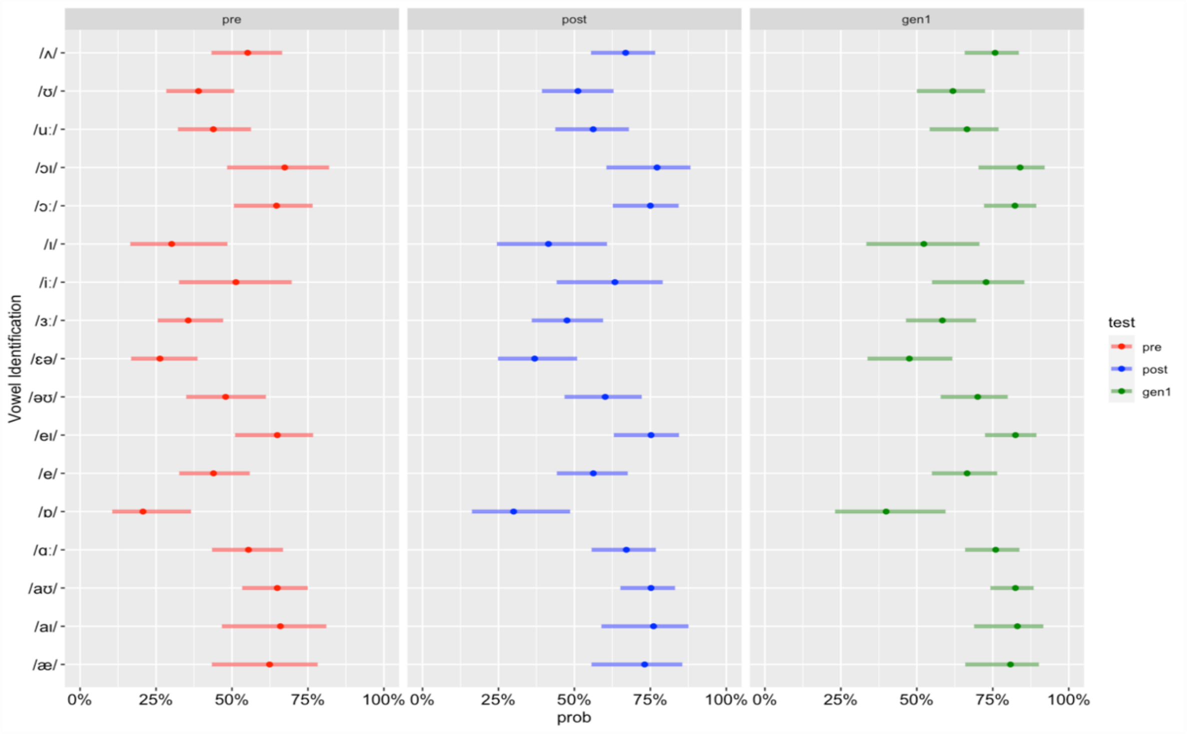


Figure 4.15 Probability of vowels considering the gen1 in relation to pre-and post-tests for the ID task.

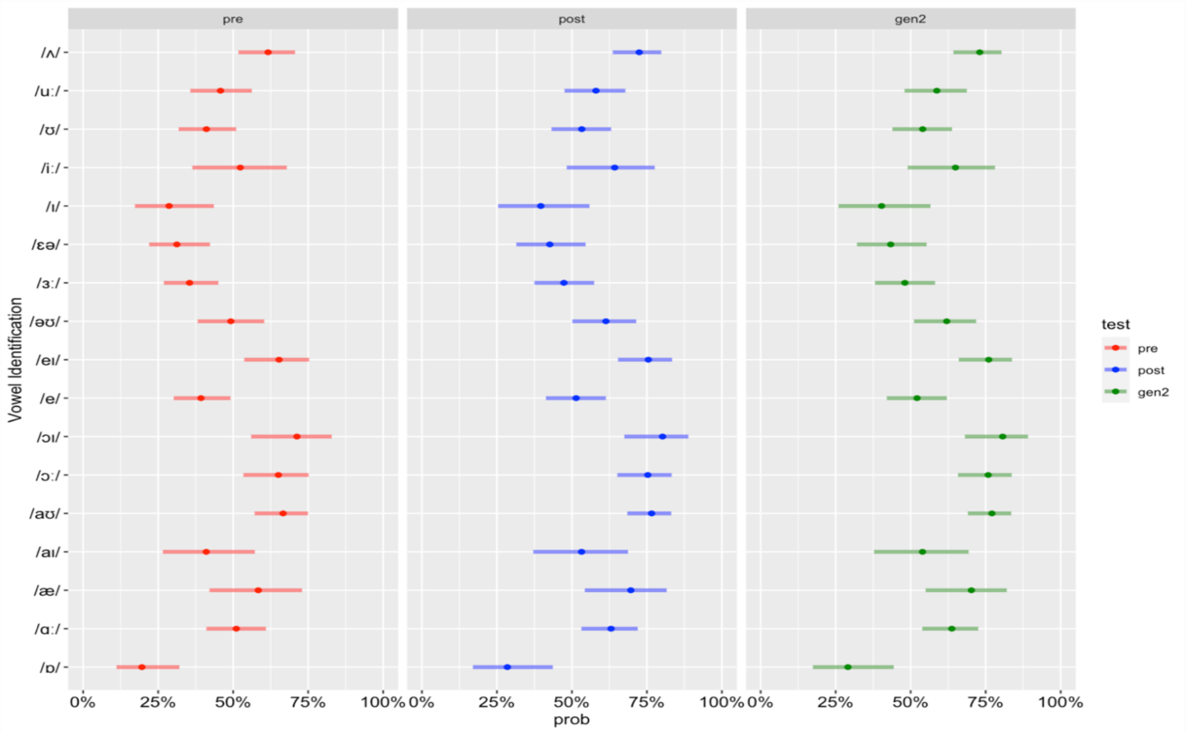


Figure 4.16 Probability of vowels considering the gen2 in relation to pre-and post-tests for the ID task.

4.4.2.4 Summary

The below points summarise the findings of the generalisation tests for the vowel identification task:

- In the gen1 test, groups A, B, and C demonstrated increased likelihoods of correct responses compared to their performance in the pre- and post-tests.
- In the gen2 test, all groups showed improvements compared to the pre-test, with results aligning closely with those seen in the post-test.
- The performance of groups A, B, and C consistently shows similarity in the gen1 and gen2 tests. These outcomes suggest that employing multiple varieties in training (groups B and C) improved the perception of other varieties heard in the generalisation tests to the same extent as a training approach focused on a single L1 variety (group A). Therefore, the inclusion of multiple accents was advantageous, as it did not impede the progress of beginners.
- Regardless of classifying vowels as easy or difficult based on learners' performance in the pre-test (above or below 50% threshold probability), the majority of vowels in the identification task scored above 50% in both gen1 and gen2 tests. This indicates the learners' capacity to adapt to the task and retain their learning.

4.4.3 Auditory Discrimination

4.4.3.1 Tests

The model analysis revealed a statistically significant improvement in response accuracy during the gen1 test ($p < 0.01$) and post-test ($p < 0.0001$) as compared to the pre-test. Specifically, with all other variables held constant, the gen1 test and post-test demonstrated probability rates of 73% and 75%, respectively. In a comparable manner, the model demonstrated a statistically significant improvement in response accuracy during both the gen2 test and post-test ($p < 0.0001$) in relation to the pre-test. The findings from the gen2 test revealed a probability rate of 78%, which exceeded the score of 73% achieved in the post-test. These results suggest that the gen2 test demonstrated a greater probability in comparison to both the gen1 test and post-test.

Table 4.33 shows a pairwise comparison analysis of the test, focusing on the performance of gen1 relative to the pre and post-tests. It can be observed that gen1 demonstrated a significantly higher level of accuracy than the pre-test (odds ratio= 0.366, $p < 0.01$), but not significantly different from the post-test (odds ratio = 1.088, $p > 0.05$). On the other hand, Table 4.34 displays a comparative assessment of gen2 in relation to the pre-and post-tests. The data demonstrate that gen2 achieved a considerably greater degree of accuracy compared to the pre-test (odds ratio = 0.281, $p < 0.001$). While the gen2 test presented higher results relative to the post-test (with an odds ratio of 0.772), no statistically significant distinctions were seen between the two ($p > 0.05$).

contrast	*Odds.ratio*	*SE*	*z.ratio*	*p.value*
pre / gen1	0.366	0.118	-3.130	0.0026
post/ gen1	1.088	0.290	0.267	0.7892

Table 4.33 Main effect of tests (pre-, post-, gen1) for the AD task, without considering interactions. The mid-point of the odds ratio is 1.

Contrast	*Odds.ratio*	*SE*	*z.ratio*	*p.value*
pre / gen2	0.281	0.0937	-3.808	0.0004
post/ gen2	0.772	0.4572	-0.733	0.4633

Table 4.34 Main effect of tests (pre-, post-, gen2) for the ID task, without considering interactions. The mid-point of the odds ratio is 1.

Figure 4.17 shows the EMMeans for the gen1 test compared to the pre- and post-tests for groups A, B, and C, while Figure 4.18 presents the EMMeans for the gen2 test. Confidence intervals indicate greater variability in the pre-test than in the gen1, gen2, and post-tests. Despite ceiling effects in the pre-test, accuracy in gen1 and gen2 was higher and comparable to the post-test, indicating effective skill application. Improved perceptual abilities likely aided in discriminating unfamiliar words and speakers.

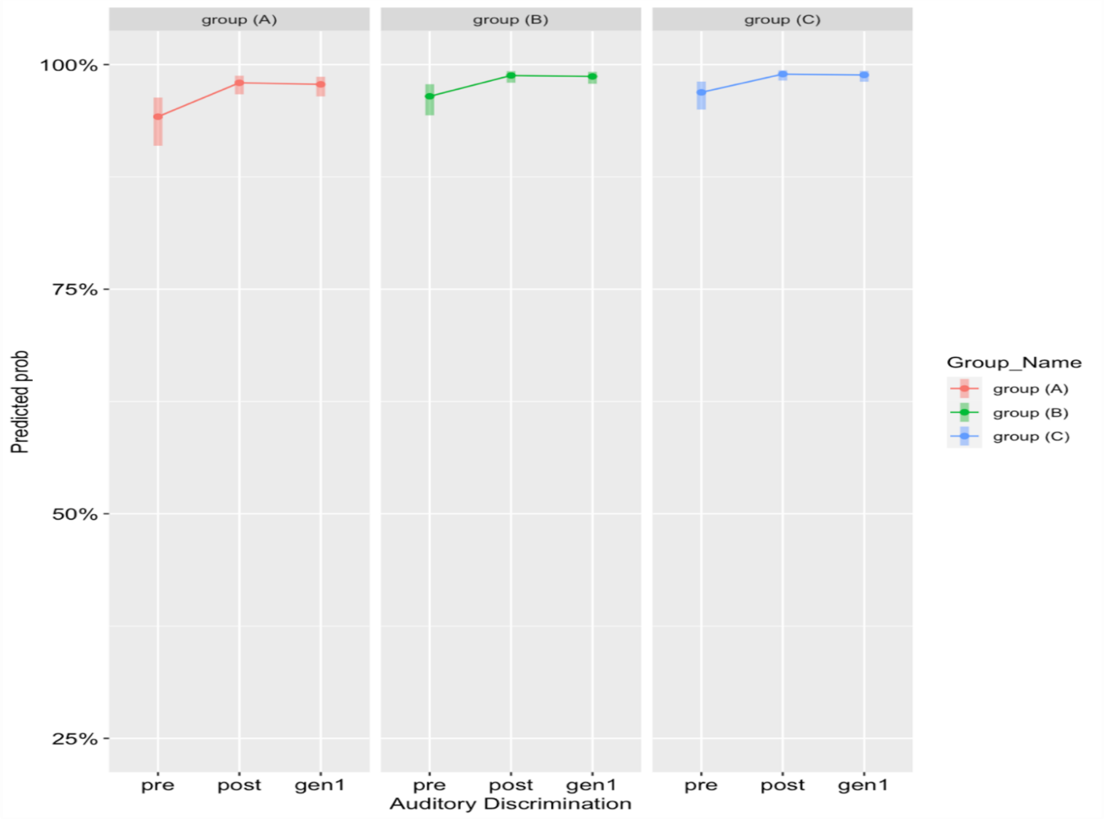


Figure 4.17 The EMMeans of gen1 relative to pre- and post-tests across groups for the AD task.

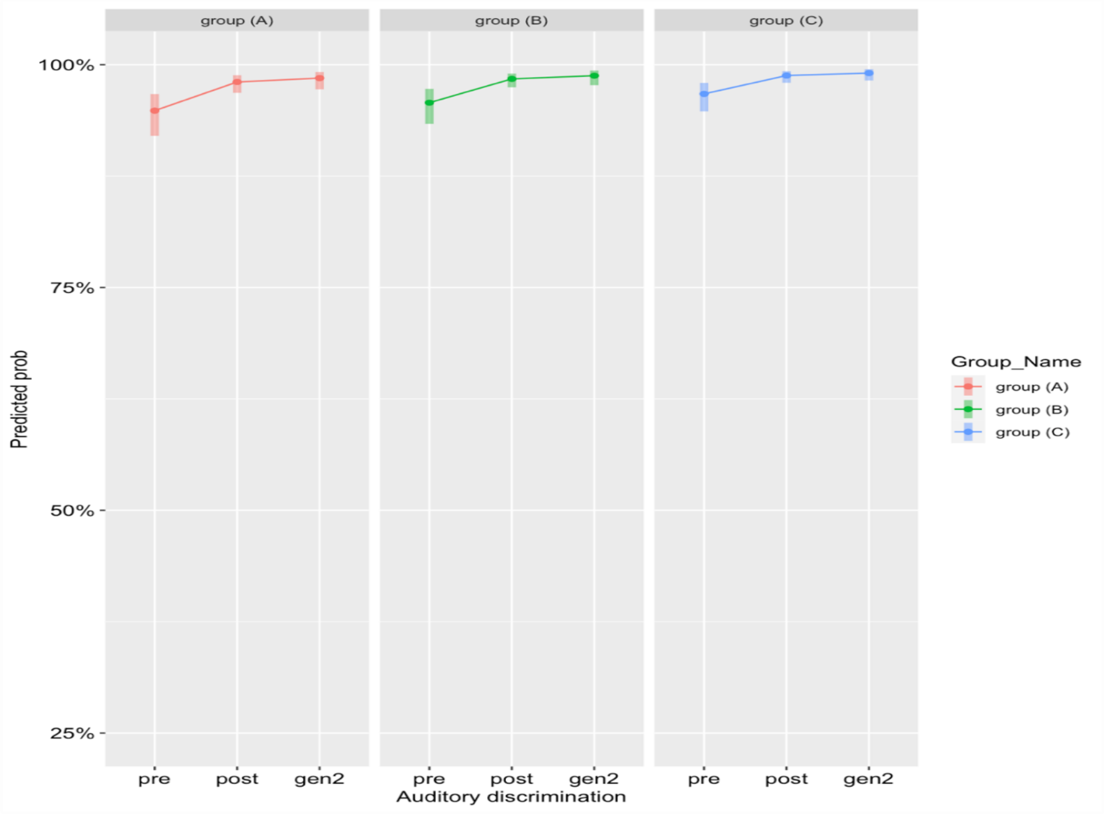


Figure 4.18 The EMMeans of gen2 relative to pre- and post-tests across groups for the AD.

The interaction results of tests across the groups for the AD task (see Tables 4.35 and 4.36) indicate that the performance levels of groups A, B, and C during gen1 and gen2 were substantially (*p* < 0.0001) higher than those during the pre-test, with improvements ranging from 2% to 4% for gen1 and between 3% and 4% for gen2 (see Tables 4.23 and 4.24 below). Alternatively, the performance of groups in gen1 and gen2 indicated a level of similarity to their performance in the post-test[63].

[63] Due to the use of raw data, the test contrasts (post – gen1 and post – gen2) revealed a high degree of statistical significance across the three groups. However, the important evaluation lies in deciding the extent of the mean difference.

Vowels	***Contrast***	***Mean Difference***	***p-value***
A	pre - post	-4%	.000
	pre - gen.1	-4%	.000
	post- gen.1	-0.01%	0.96
B	pre - post	-3%	.000
	pre - gen.1	-3%	.000
	post- gen.1	-0.01%	0.94
C	pre - post	-2%	.000
	pre - gen.1	-2%	.000
	post- gen.1	-0.01%	.000

Table 4.35 Interactions of tests (gen1, pre-, post-) across the groups for the AD task.

Vowels	***Contrast***	***Mean Difference***	***p-value***
A	pre - post	-4%	.000
	pre - gen.2	-4%	.000
	post- gen.2	-0.6%	.000
B	pre - post	-3%	.000
	pre - gen.2	-4%	.000
	post- gen.2	-0.5%	.000
C	pre - post	-2%	.000
	pre - gen.2	-3%	.000
	post- gen.2	-0.4%	.000

Table 4.36 Interactions of tests (gen2, pre-, post-) across the groups for the AD task.

4.4.3.2 Groups

The coefficients obtained from the mixed-effect models offered valuable insights into the effect of gen1 and gen2 tests on group performance in the AD task while considering the impact of both pre-and post-test results. Upon inclusion of the gen1, pre-, post-tests, it was seen that groups B and C showcased a considerably greater level of accuracy in giving accurate responses (63%, p < 0.01, 66%, p < 0.0001) respectively, outperforming the performance of the reference group A. In contrast, with the incorporation of the gen2, pre- and post-test, it was seen that group C tended to show significance with a probability of 62% (p = 0.051) of providing accurate responses compared to group A, whereas the accuracy of group B did not display statistical significance (55%, p < 0.05).

Table 4.37 presents the results of the pairwise comparison analysis, with a particular focus on evaluating the performance of groups with respect to the model that includes gen1, pre-and post-tests. The findings indicated a significant statistical difference between groups A and B (odds ratio = 0.595, *p* < 0.01) as well as between groups A and C (odds ratio = 0.518, *p* < 0.01), with groups B and C demonstrating higher scores compared to group A. However, no statistically significant difference was seen between groups B and C (odds ratio = 0.871, *p* > 0.05).

On the other hand, the findings of the pairwise comparison analysis, including gen2 alongside with pre-and post-tests, were shown in Table 4.38. The results indicate that group B demonstrated greater success compared to group A, with an odds ratio of 0.819. Additionally, group C showed even higher performance than both group A (odds ratio = 0.623) and group B (odds ratio = 0.761). Nevertheless, the observed performance disparities did not reach statistical significance (*p* > 0.05), suggesting that the overall group results were comparable.

Contrast	***Odds.ratio***	***SE***	***z.ratio***	***p.value***
group (A) / group (B)	0.595	0.113	-2.731	0.0095
group (A) / group (C)	0.518	0.102	-3.329	0.0026
group (B) / group (C)	0.871	0.176	-0.683	0.4944

Table 4.37 The main effect of the groups in the AD task over the pre-, post-, and gen1-tests, without considering interactions. The odds ratio's midpoint stands at 1.

Contrast	***Odds.ratio***	***SE***	***z.ratio***	***p.value***
group (A) / group (B)	0.819	0.193	-0.850	0.3956
group (A) / group (C)	0.623	0.151	-1.947	0.1546
group (B) / group (C)	0.761	0.184	-1.129	0.3887

Table 4.38 The main effect of the groups in the AD task over the pre-, post-, and gen2-tests, without considering interactions. The odds ratio's midpoint stands at 1.

Figures 4.19 and 4.20 show the EMMeans for average group performance on the AD task across gen1 and gen2 tests, considering pre- and post-test results. The CIs indicate higher variability during the pre-test for groups A, B, and C, with reduced variability in the generalisation and post-tests. Despite initial ceiling effects, slight improvement was observed during the generalisation and post-tests.

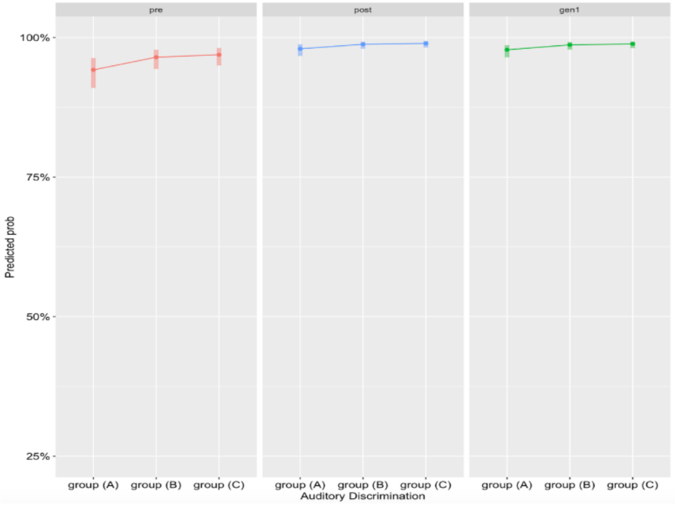


Figure 4.19 The EMMeans of the groups considering the gen1 in relation to pre-and post-test for the AD task (The threshold is 50%).

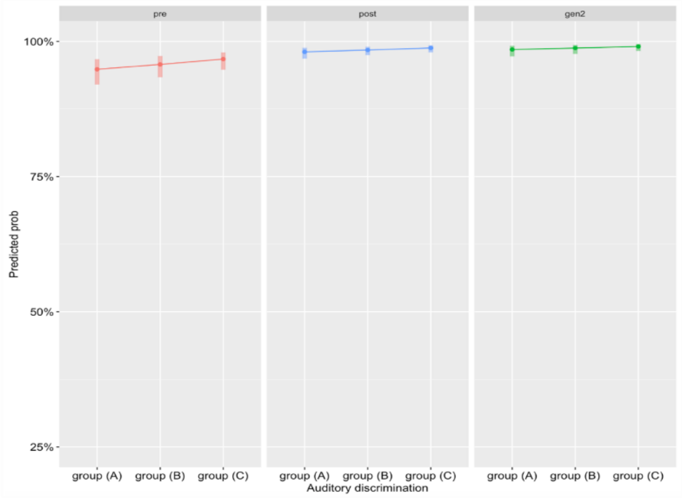


Figure 4.20 The EMMeans of the groups considering the gen2 in relation to pre-and post-test for the AD task (The threshold is 50%).

The multiple comparison t-test was performed to determine the interactions of groups across the tests (gen1, gen2) comparing their performance in the pre-and post-test (see Table 4.39 and 4.40[64]). Throughout the gen1 and gen2 tests, all three groups demonstrated similar levels of performance, with group C holding a marginal edge over groups A and B (0.2% to 1%), and group B slightly outperforming group A (0.3% to 1%). This marginal improvements in the performances of groups B and C were roughly mirrored in the pre- and post-tests. Hence, it is evident that the disparities among the three groups are negligible, indicating that their performances closely resemble each other.

Test	***Contrast (groups)***	***Mean Difference***	***p-value***
Gen1	A - B	-1%	0.000
	A - C	-1%	0.000
	B - C	-0.2%	0.000
Pre-test	A - B	-1%	0.000
	A - C	-1%	0.000
	B - C	-0.2%	0.000
Post-test	A - B	-3%	0.000
	A - C	-3%	0.000
	B - C	-0.5%	0.000

Table 4.39 Interactions of groups across tests (gen1, pre-, post-) for the AD task.

Test	***Contrast (groups)***	***Mean Difference***	***p-value***
Gen2	A - B	-0.3%	0.000
	A - C	-1%	0.000
	B - C	-0.3%	0.000
Pre-test	A - B	-1%	0.000
	A - C	-2%	0.000
	B - C	-1%	0.000
Post-test	A - B	-0.4%	0.000
	A - C	-1%	0.000
	B - C	-0.43%	0.000

Table 4.40 Interactions of groups across tests (gen2, pre-, post-) for the AD task.

[64] Due to the use of raw data, the group comparisons across all tests revealed a high degree of statistical significance. Yet, the critical assessment lies in determining the extent of the mean difference.

4.4.3.3 Vowels

The multiple comparison t-test was performed to determine the interactions of tests (gen1, gen2) in relation to the pre-and post-test across the vowels for the AD task (See Tables G. VI and G.VII in Appendix F). The results showed a significant improvement in vowel performance during gen1, with an increase ranging from 1% to 6% ($p <$ 0.0001), and during gen2, with an increase from 1% to 7% ($p <$ 0.0001), when compared to the pre-test. Meanwhile, no substantial disparities were observed among the vowels when comparing the two generalisation tests to the post-test.

Tables 4.41 and 4.42 show how vowels impact response accuracy in gen1 and gen2 compared to the AD task's pre-and post-test (without interactions). The accuracy level in distinguishing vowels during gen1 and gen2 aligned closely with the results of the post-test. Although the pre-test revealed great accuracy for all vowels (with all scores above 50%), the gen1 and gen2 tests showed even higher accuracy. For instance, in the gen1 test, the accuracy of the vowels /e/, /ʌ/, /ɛə/, /ʊ/, /ɑː/, /aɪ/, and /əʊ/ experienced increments: there was a 6% increase for /e/, 4% for /ʌ/ and /ɛə/, 3% for /ʊ/ and /ɑː/, and 2% for /aɪ/ and /əʊ/. Similarly, in the gen2 test compared to the pre-test, there was a 7% rise for /e/, 5% for /ɛə/, 3% for /ʌ/, /ʊ/, and /ɑː/, 2% for /aɪ/ and /əʊ/.

In short, the vowel accuracy in both the gen1 and gen2 tests of the auditory discrimination task exceeded the pre-test results while demonstrating a similar pattern to the post-test results.

Vowels	Pre-test	Post-test	Gen1
/aʊ/	97.7%	99.2%	99.2%
/ɔɪ/	97.4%	99.1%	99.0%
/əʊ/	97.4%	99.1%	99.0%
/iː/	97.4%	99.1%	99.0%
/aɪ/	97.3%	99.0%	99.0%
/æ/	97.0%	99.0%	98.9%
/eɪ/	96.9%	98.9%	98.9%
/ɪ/	96.9%	98.9%	98.8%
/ɜː/	95.5%	98.4%	98.3%
/ʊ/	95.4%	98.4%	98.3%
/uː/	95.2%	98.3%	98.2%
/ɑː/	94.5%	98.1%	97.9%
/ɛə/	94.0%	97.9%	97.7%
/ʌ/	93.7%	97.8%	97.6%
/ɔː/	93.6%	97.8%	97.6%
/ɒ/	98.6%	98.7%	98.6%
/e/	90.2%	96.5%	96.2%

Table 4.41 The EMMeans of vowels considering the gen1 in relation to pre- and post-test for the AD task.

Vowels	Pre-test	Post-test	Gen 2
/ɔɪ/	98.2%	99.3%	99.5%
/aʊ/	97.4%	99.0%	99.3%
/əʊ/	97.3%	99.0%	99.2%
/ɪ/	97.0%	98.9%	99.1%
/iː/	96.9%	98.9%	99.1%
/aɪ/	96.8%	98.8%	99.1%
/æ/	96.6%	98.7%	99.0%

/ɒ/	96.3%	98.6%	98.9%
/ʌ/	95.8%	98.4%	98.8%
/ɑː/	95.3%	98.2%	98.6%
/ʊ/	95.3%	98.3%	98.6%
/eɪ/	95.1%	98.2%	98.6%
/ɔː/	94.2%	97.8%	98.3%
/uː/	94.2%	97.8%	98.3%
/ɜː/	94.0%	97.7%	98.2%
/ɛə/	92.9%	97.3%	97.9%
/e/	89.7%	96.0%	96.9%

Table 4.42 The EMMeans of vowels considering the gen2 in relation to pre-and post-test for the AD task[65].

Figures 4.21 and 4.22 provide visual presentations of EMMeans of vowel responses during the AD task of gen1 and gen2 evaluations, taking into consideration the results of both pre and post-tests. It appears that the average predicted probability of accurate responses in the pre-test is typically less than what's seen in the gen1, gen2, and post-tests. However, vowel accuracy on both the gen1 and gen2 tests was comparable to the performance on the post-test. The two charts also highlight that the vowels /ɛə/, /e/, and /ɔː/ demonstrate expanded CIs, suggesting a higher degree of variability around their mean predicted probabilities. This hints at a degree of inconsistency in the participants' responses, which could be responsible for the less reliable mean predicted probabilities for these vowels. On the other hand, the remaining vowels display narrow CIs, representing minor variability around their average predicted probability, hinting at more precise average predicted probabilities.

[65] The correct response probabilities are represented by the following colours: (dark green) for high, (medium green) for moderate, and (pale green) for low.

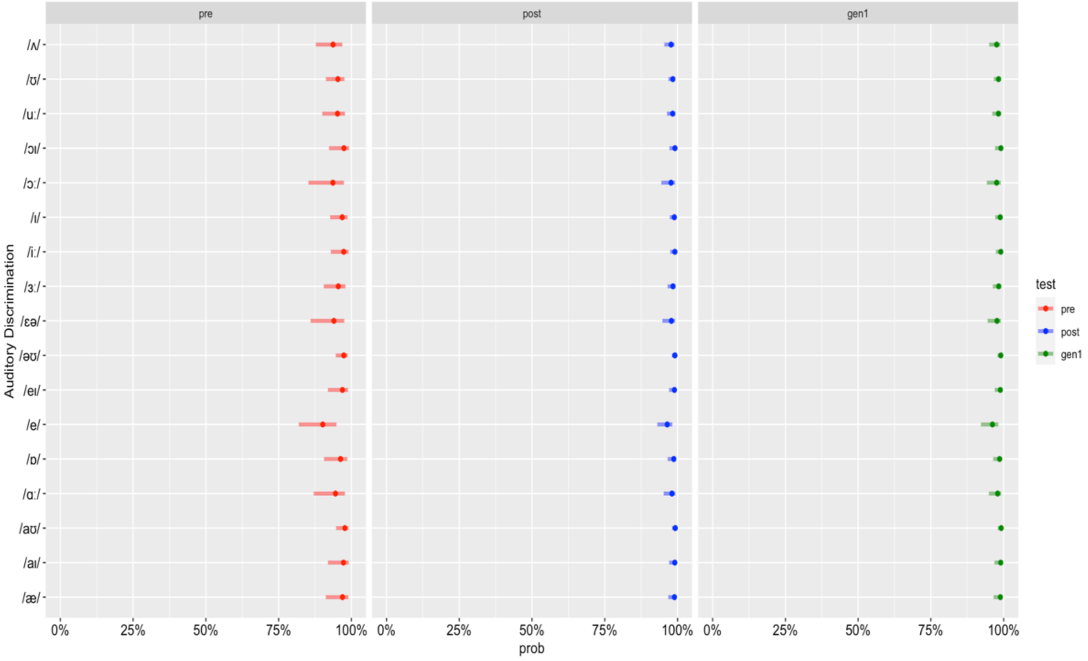


Figure 4.21 The probability of vowels considering the gen1 in relation to pre-and post-test **for** the AD task (The threshold is 50%).

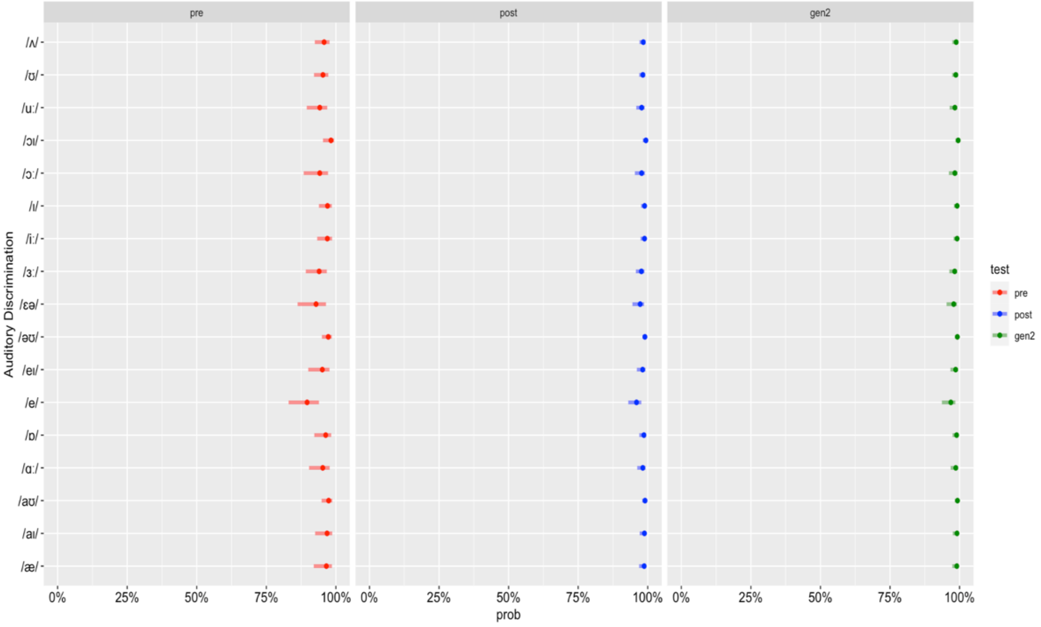


Figure 4. **22** The probability of vowels considering the gen2 in relation to pre-and post-test for the AD task (The threshold is 50%).

4.4.3.4 Summary

The below points summarise the findings of the generalisation tests for AD task:

- In the gen1 and gen2 tests, it was seen that all vowels showed significant improvement in comparison to the pre-test, and these improvements were found to be similar to those observed in the post-test. This demonstrates the learners' capacity to adapt to the task and maintain their learning.

- Groups A, B, and C demonstrate similar levels of performance in the gen1 and gen2 tests while engaging in the AD task. These findings suggest the use of multiple L1 varieties with a single L2 variety (training group C) did not provide a significant improvement in vowel perception compared to the use of multiple L1 varieties (training group B) and a single L1 variety (group A) during the AD task. This finding aligns with the outcomes of the identification task.

- The capacity to discriminate between vowels seemed to be very straightforward in the pre-test during the AD task, but it became much easier in the post-test and the generalisation tests, so explaining the efficacy of the training.

4.4.4 Category Discrimination

4.4.4.1 Tests

The analysis of the model results showed a significant improvement in the probability of accurate responses for both the gen1 test and post-test (p < 0.0001) as compared to the pre-test baseline. The probabilities for accurate responses were 69% and 73%, respectively. A parallel pattern was observed in gen2 (68%, p < 0.0001) and the post-test (73%, p < 0.0001) when compared to the pre-test.

Tables 4.43 and 4.44 provide the results of the paired comparison test, illustrating the accuracy of vowels during gen1 and gen2 in relation to both pre- and post-tests. The results indicate that there was no statistically significant difference in accuracy between the gen1 and post-tests (odds ratio=1.246, p > 0.05) and between gen2 and post-tests (odds ratio=1.273, p > 0.05). However, it was clear that both gen1 and gen2 showed a significantly higher probability of providing correct answers in comparison to the pre-test, with an odds ratio of 0. 455 (p < 0.001) and 0. 465 (p < 0.001), respectively.

Contrast	*Odds.ratio*	*SE*	*z.ratio*	*p.value*
pre / gen1	0.455	0.0933	-3.842	0.0002
post / gen1	1.246	0.1629	1.085	0.2779

Table 4.43 The main effect of tests (pre-, post-, gen1) related to the CD task, without considering interactions. The mid-point of the odds ratio is 1.

Contrast	*Odds.ratio*	*SE*	*z.ratio*	*p.value*
pre / gen2	0.465	0.0988	-3.606	0.0005
post/ gen2	1.273	0.1669	1.136	0.2559

Table 4.44 The main effect of tests (pre-, post-, gen2) related to the CD task, without considering interactions. The mid-point of the odds ratio is 1.

Visual representations of EMMeans for the gen1 and gen2 tests are shown in Figure 4.23 and Figure 4.24, respectively. For each test, comparisons are made against the pre-and post-tests for each of the groups (A, B, and C) participating in the CD task. The consistent width of the confidence intervals, which represent the average predicted probabilities for both graphs, indicates a substantial level of reliability in predicting average test results across all groups.

Figure 4.23 illustrates that for the gen1 test, the average probability of scoring correct answers across the groups ranges between 54% and 61%. Similarly, Figure 4.24 presents the gen2 test results, with an average correct answer probability ranging from 52% to 61%. This indicates that performances on the two generalisation tests were approximately equal. The accuracy of responses for all groups during the two generalisation tests was higher than the pre-test range of 34% to 42%, although it still fell short of the post-test range of 58% to 66%. As discussed in section 4.4.1, generalisation occurs when scores are as high as or higher than the post-test scores and differ from the pre-test scores (Carlet & Cebrian, 2019). Given that the performance of all groups showed only slight deviation from the post-test scores while displaying a greater difference from the pre-test results, generalisation appears feasible in the CD task. This suggests some degree of generalisation to new words and items, demonstrating the effective transfer of learned abilities.

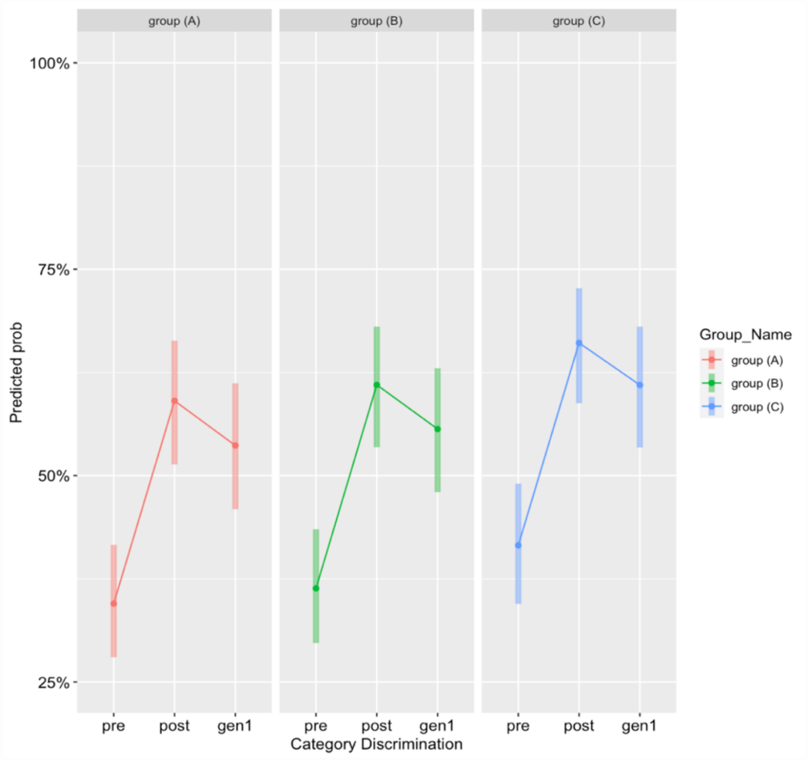


Figure 4.23 The EMMeans of gen1 relative to pre- and post-tests across the groups for the CD task (The threshold is 50%).

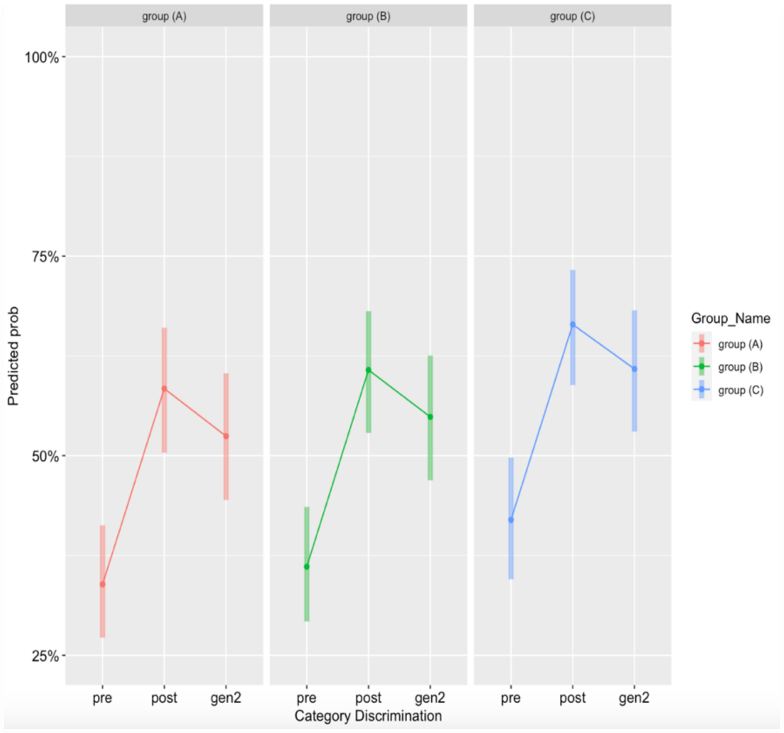


Figure 4.24 The EMMeans of gen2 relative to pre- and post-tests across the groups for the CD (The threshold is 50%).

According to the interactions of tests across the groups (see Tables 4.45 and 4.46), the three groups performed substantially better on the gen1 and gen2 tests compared to the pre-test by 20% to 21% ($p < 0.0001$) and by 21% to 22% ($p < 0.0001$), respectively. In contrast, the levels of accuracy demonstrated by participants in gen1 and gen2 were lower when compared to the post-test results, with a disparity of only 3% seen for each respective group.

Groups	***Contrast***	***Mean Difference***	***p-value***
A	pre - post	-24%	0.000
	pre - gen1	-20%	0.000
	post- gen1	3%	0.000
B	pre - post	-24%	0.000
	pre - gen1	-21%	0.000
	post- gen1	3%	0.000
C	pre - post	-24%	0.000
	pre - gen1	-21%	0.000
	post- gen1	3%	0.000

Table 4.45 Interactions of tests (gen1, pre-, post-) across the groups for the CD task.

Groups	***Contrast***	***Mean Difference***	***p-value***
A	pre – post	-24%	0.000
	pre – gen.2	-21%	0.000
	post- gen.2	3%	0.000
B	pre – post	-24%	0.000
	pre – gen.2	-22%	0.000
	post- gen.2	3%	0.000
C	pre – post	-24%	0.000
	pre – gen.2	-22%	0.000
	post- gen.2	3%	0.000

Table 4.46 Interactions of tests (gen2, pre-, post-) across the groups for the CD task.

4.4.4.2 Groups

The coefficients obtained from the mixed-effect models provided valuable insights into the influence of gen1 and gen2 on group performance in the CD task, all while accounting for both pre and post-test outcomes. Upon the inclusion of the gen1, pre- and post-test, it appeared that group C showed a noteworthy response accuracy of

57% (p= 0.006), exceeding the performance of group A (the reference category). In contrast, group B's response rate of 52% did not display any statistically significant increase (*p* > 0.05). The observed pattern demonstrated consistency when the gen1, pre-and post-tests were included. Group C had a statistically significant accuracy rate of 59% (*p* < 0.01), while group B showed a somewhat average accuracy rate of 52% (*p* > 0.05).

Table 4.47 provides a comprehensive analysis of the pairwise comparison test, specifically examining the overall performance of groups, including gen1, pre-and post-tests. Group C had a considerably higher performance level than group A (odds ratio= 0.741, *p* < 0.05). Additionally, there was a tendency for group C to outperform group B (odds ratio= 0.803, p= 0.0584). Nevertheless, the analysis yielded no statistically significant difference between groups A and B (odds ratio = 0.923, *p* > 0.05).

Contrast	***Odds.ratio***	***SE***	***z.ratio***	***p.value***
group (A) / group (B)	0.923	0.0988	-0.753	0.4517
group (A) / group (C)	0.741	0.0809	-2.747	0.0180
group (B) / group (C)	0.803	0.0853	-2.065	0.0584

Table 4.47 The main effect of the groups in the CD task over the pre-, post-, and gen1-tests, without considering interactions. The odds ratio's midpoint stands at 1.

On the other hand, Table 4.48 presents a thorough analysis of the pairwise comparison test, particularly focusing on the overall performance of groups with the inclusion of gen2, pre-and post-tests for the CD task. There was a statistically significant difference between groups A and C (odds ratio = 0.709, *p* < 0.01) and between groups B and C (odds ratio = 0.781, *p* < 0.05), with group C scoring higher than both groups A and B. Yet, no significant difference was seen between groups A and B (odds ratio = 0.907, *p* > 0.05).

Contrast	Odds.ratio	SE	z.ratio	p.value
group (A) / group (B)	0.907	0.0993	-0.888	0.3743
group (A) / group (C)	0.709	0.0790	-3.085	0.0061
group (B) / group (C)	0.781	0.0851	-2.266	0.0352

Table 4.48 The main effect of the groups in the AD task over the pre-, post-, and gen2-tests, without considering interactions. The odds ratio's midpoint stands at 1.

Figures 4.25 and 4.26 illustrate the EMMeans, illustrating the probability of accurate answers from each group during the gen1 and gen2 tests compared to the pre-test and post-test results. The confidence intervals for the predicted average probabilities for groups A, B, and C show stability across all evaluations, suggesting a uniform accuracy in predicting each group's average scores.

Upon analysing the test results of the gen1 test, it was seen that group C had a greater likelihood of providing accurate answers, with a percentage of 60.9%, in contrast to groups B and A, which achieved 55.6% and 53.6%, respectively. The same pattern was seen in the gen2 test when group C once again had a higher likelihood of providing accurate replies (60.8%) compared to groups B (54.8%) and A (52.4%). Although group C performed better than groups A and B in the gen1 and gen2 tests, these findings do not provide evidence of any subsequent advancement. Instead, they only represent a minor advantage in performance for group C seen during the pre-test. Hence, the performance shown by all three groups during the generalisation assessments of the CD task is comparable.

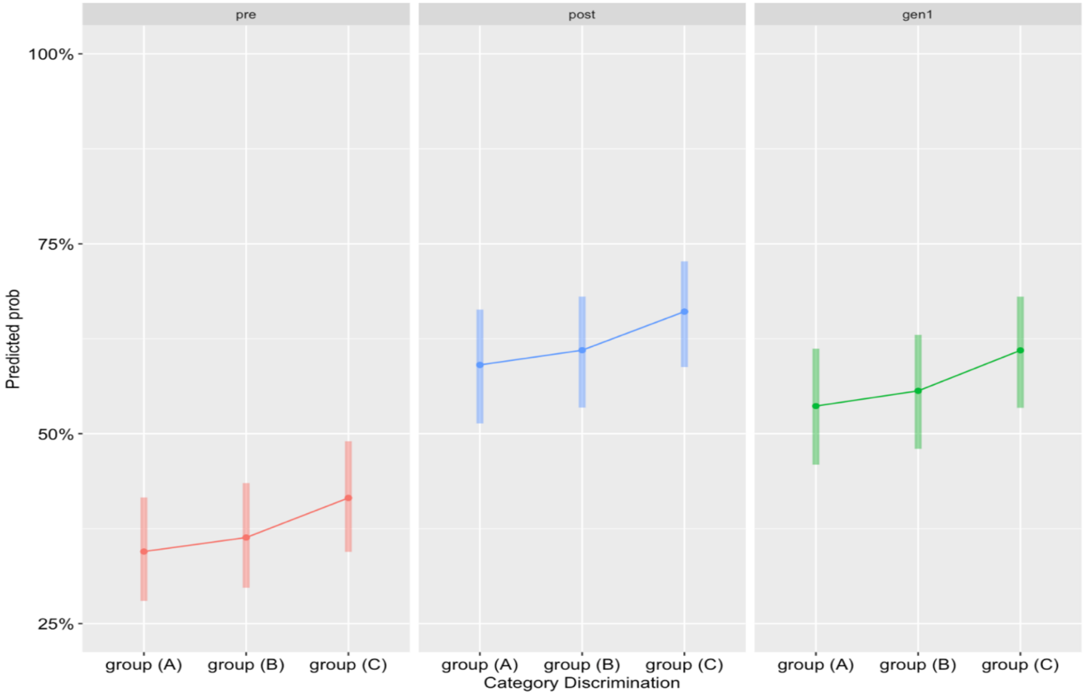


Figure 4.25 The EMMeans of the groups considering the gen1 in relation to pre-and post-test for the CD task (The threshold is 50%).

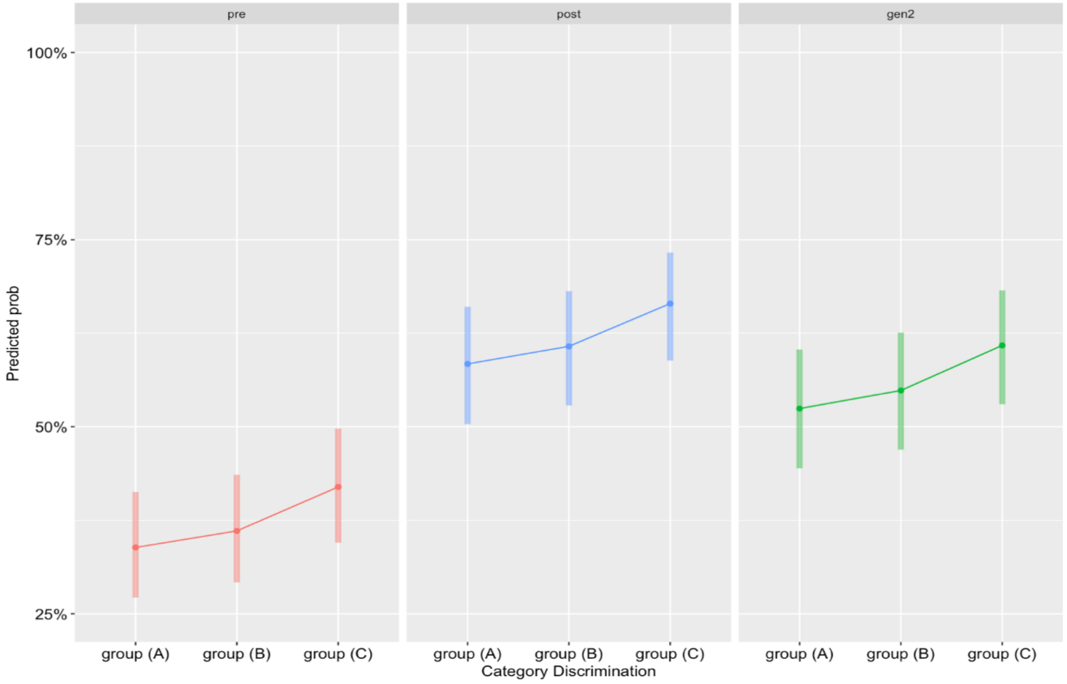


Figure 4.26 The EMMeans of the groups considering the gen2 in relation to pre-and post-test for the CD task (The threshold is 50%).

The multiple comparison t-test was used to examine the interactions of groups across the tests (gen1, gen2) in relation to their performance on the pre-and post-test for the CD task (see Tables 4.49, 4.50 below). Group B had slightly better results compared to group A by 2% during the gen1 and gen test. On the other hand, group C significantly had higher performance than both group A and by (by 5% to 8%, *p* < 0.0001). The observed improvement between the groups in gen1 and gen2 was similarly seen in the pre-and post-tests. Hence, the performance demonstrated by the groups in the two generalisation tests of the CD task was equivalent, and the observed variations between the groups can be attributed to their respective performance during pre-and post-tests.

Test	***Contrast (groups)***	***Mean Difference***	***p-value***
Gen1	A - B	-2%	0.000
	A - C	-7%	0.000
	B - C	-5%	0.000
Pre-test	A - B	-2%	0.000
	A - C	-7%	0.000
	B - C	-5%	0.000
Post-test	A - B	-2%	0.000
	A - C	-7%	0.000
	B - C	-5%	0.000

Table 4.49 Interactions of groups across tests (gen1, pre-, post-) for the CD task.

Test	***Contrast (groups)***	***Mean Difference***	***p-value***
Gen2	A - B	-2%	0.000
	A - C	-8%	0.000
	B - C	-6%	0.000
Pre-test	A - B	-2%	0.000
	A - C	-8%	0.000
	B - C	-6%	0.000
Post-test	A - B	-2%	0.000
	A - C	-8%	0.000
	B - C	-6%	0.000

Table 4.50 Interactions of groups across tests (gen2, pre-, post-) for the CD task.

4.4.4.3 Vowels

Within the context of the CD task, gen1 and gen2 tests made use of the multiple comparison t-test (Refer to Tables G.VIII and G.IX in Appendix F) to show the interactions between tests (how the generalisation tests interacted with the pre-and post-tests) across vowels. The results showcased significant improvements for each vowel in both gen1 (14% to 19%, $p < 0.0001$) and gen2 (16% to 19%, $p < 0.0001$), as compared to the pre-test. Compared to the post-test, on the other hand, there was a notable increase in vowel accuracy during the gen1 test (by 3% to 5%, $p < 0.0001$) and gen 2 (by 4% to 6%, $p < 0.0001$).

The individual impact of vowels on response accuracy during the gen1 relative to the pre-and post-tests is presented in Table 4.51 (without interactions**)**. The accuracy of all vowels in the gen1 test demonstrated lower values in comparison to those observed in the post-test. In a detailed examination, it was observed that there was a 3% decrease in the accuracy of the vowel /ɑː/, a 4% decrease for the vowels /aʊ/ and /e/, and a 5% decrease for the majority of the remaining vowels (/aɪ/, /æ/, /ɔː/, /ɔɪ/, /eɪ/, /iː/, /ʌ/, /ɒ/, /ɜː/, /ɛə/, /ɪ/, /ʊ/, /uː/, /əʊ/). By contrast, the accuracy of all vowels in the gen1 test exhibited higher values in comparison to those observed in the pre-test. While the vowels /ɑː/, /aʊ/, and /e/ demonstrated a high level of accuracy at the gen1 test and pre-test (exceeding 50%), the accuracy increased during the gen1 by 14% for /ɑː/, 17% for /aʊ/ and 19% for /e/. In the gen1 test, the vowels /ɑː/, /aʊ/, /e/, /iː/, /ʊ/, /ɔɪ/, /ʌ/, /ɒ/, /uː/, /əʊ/, and /ɛə/ had a probability rate greater than 50%, while they achieved a probability rate of less than 50% during the pre-test. For instance, the vowel /əʊ/ showed an accuracy rate of 53% in the gen1 test, which indicates a 19% increase in accurate responses compared to the pre-test's accuracy rate of 34%. Although the vowels /ɪ/, /eɪ/, /aɪ/, /ɔː/, /æ/, and /ɜː/ showed performance level below 50% in both gen1 test and pre-test, an improvement in accuracy was seen during the gen1 test. More precisely, there was a 19% increase for /ɪ/, an 18% increase for /eɪ/, /aɪ/, and /ɔː/, and a 17% increase for /æ/ and /ɜː/ when compared to the pre-test.

Table 4.52, on the other hand, demonstrates the individual impact of vowels on response accuracy during the gen2 relative to the pre-and post-tests, without taking interactions into account. The accuracy of all vowels in the gen2 test demonstrated lower values in comparison to those observed in the post-test. During the gen2 test, it was noted that there was a 4% decline in the accuracy of the vowel /ɑː/, along with a 6% decrease for the remaining vowels (/aɪ/, /aʊ/, /æ/, /iː/, /ɔː/, /ɔɪ/, /eɪ/, /ʌ/, /e/, /ɜː/, /ɛə/, /ɪ/, /uː/, /əʊ/, /ʊ/, and /ɒ/). However, the accuracy of all vowels in the gen2 test was greater compared to what was observed in the pre-test. While the vowel /ɑː/ demonstrated a high level of accuracy (greater than 50%) on the pre-test, the accuracy increased by 17% on the gen2 test. The vowels /ɛə/, /eɪ/, /ɪ/, and /ɜː/ had a low probability (below 50%) in both the gen2 test as well as in the pre-test, however, the accuracy of vowels increased during the gen2 test compared to pre-test: 17% for /ɜː/, 18% for /eɪ/, and 19% for /ɛə/ and /ɪ/. Even though the vowels (/aʊ/, /æ/, /aɪ/, /ɔː/, /ɔɪ/, /iː/, /ʌ/, /ɒ/, /e/, /ʊ/, /uː/, /əʊ/ had an accuracy rate of less than 50% on the pre-test, they scored over 50% on the gen2 test. In specific, /aʊ/ and /æ/ increased by 17%, /aɪ/ and /ɔː/ rose by 18% while increased by /ɔɪ/, /iː/, /ʌ/, /ɒ/, /e/, /ʊ/, /uː/, /əʊ/ by 19%.

In brief, the vowel accuracy observed in the generalisation assessments of the category discrimination task exceeded the pre-test results. Still, it demonstrated a lower level of accuracy in comparison to the post-test outcomes.

Vowels	***Pre-test***	***Post-test***	***Gen1***
/ɑː/	69.7%	86.3%	83.5%
/aʊ/	58.0%	79.1%	75.2%
/e/	50.1%	73.4%	68.9%
/iː/	49.1%	72.6%	68.0%
/ʊ/	40.3%	64.9%	59.7%
/ɔɪ/	38.1%	62.8%	57.5%
/ʌ/	37.4%	62.1%	56.7%
/ɒ/	35.3%	59.9%	54.5%
/uː/	35.0%	59.6%	54.2%
/əʊ/	34.3%	58.8%	53.4%
/ɛə/	32.2%	56.5%	51.1%
/ɪ/	30.4%	54.5%	49.0%

/eɪ/	29.2%	53.1%	47.6%
/aɪ/	28.8%	52.6%	47.1%
/ɔː/	26.6%	49.8%	44.4%
/æ/	25.3%	48.1%	42.6%
/ɜː/	24.2%	46.7%	41.2%

Table 4.51 The EMMeans of vowels considering the gen1 in relation to pre- and post-test for the CD task[67].

Vowels	Pre-test	Post-test	Gen2
/ɑː/	60.8%	81.0%	77.0%
/iː/	43.2%	67.6%	62.1%
/ɒ/	43.2%	67.6%	62.1%
/aʊ/	43.4%	67.8%	62.3%
/ʊ/	42.3%	66.7%	61.2%
/e/	40.9%	65.5%	59.9%
/ɔɪ/	39.3%	64.0%	58.2%
/uː/	38.8%	63.5%	57.7%
/æ/	37.4%	62.1%	56.3%
/aɪ/	35.5%	60.1%	54.2%
/ʌ/	35.2%	59.8%	53.9%
/əʊ/	32.9%	57.3%	51.3%
/ɔː/	32.0%	56.4%	50.4%
/ɛə/	29.2%	53.1%	47.1%
/eɪ/	28.7%	52.5%	46.4%
/ɪ/	28.6%	52.3%	46.3%
/ɜː/	26.3%	49.4%	43.4%

Table 4.52 The EMMeans of vowels considering the gen2 in relation to pre-and post-test for the CD task.

[67] The correct response probabilities are represented by the following colours: (dark green) for high, (medium green) for moderate, and (pale green) for low. Incorrect response probabilities use (dark red) to indicate high, (medium red) to represent indicate, and (pale red) to indicate low.

Figures 4.27 and 4.28 visually depict the EMMeans of vowel responses during the gen1 and gen2 tests of the CD task, considering the outcomes of both pre and post-tests. As demonstrated, the average probability of responding correctly in the pre-test seems to be lower compared to that observed in the gen1, gen2, and post-test. However, a similar increase in accuracy can be observed from the post-tests to both gen1 and gen2 tests. The confidence intervals for the average predicted probabilities of vowels across the tests in the two charts appear to be stable. This consistency in the confidence intervals implies precise mean predicted probabilities for those vowels.

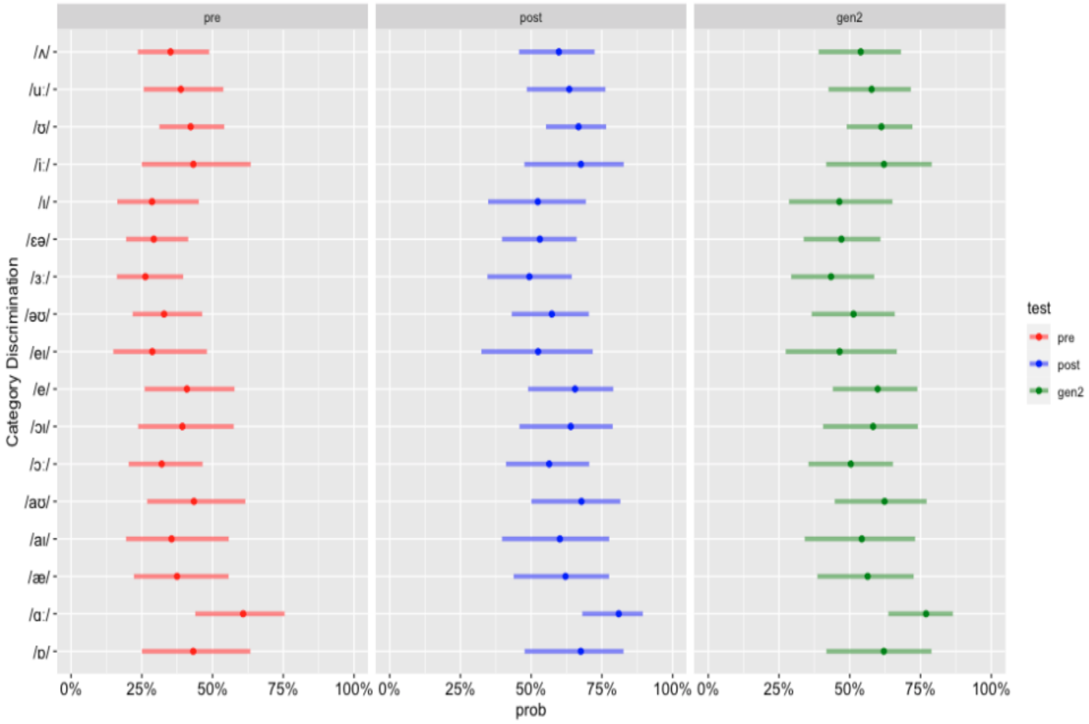


Figure 4.27 The probability of vowels considering the gen1 in relation to pre-and post-test for the CD task (The threshold is 50%).

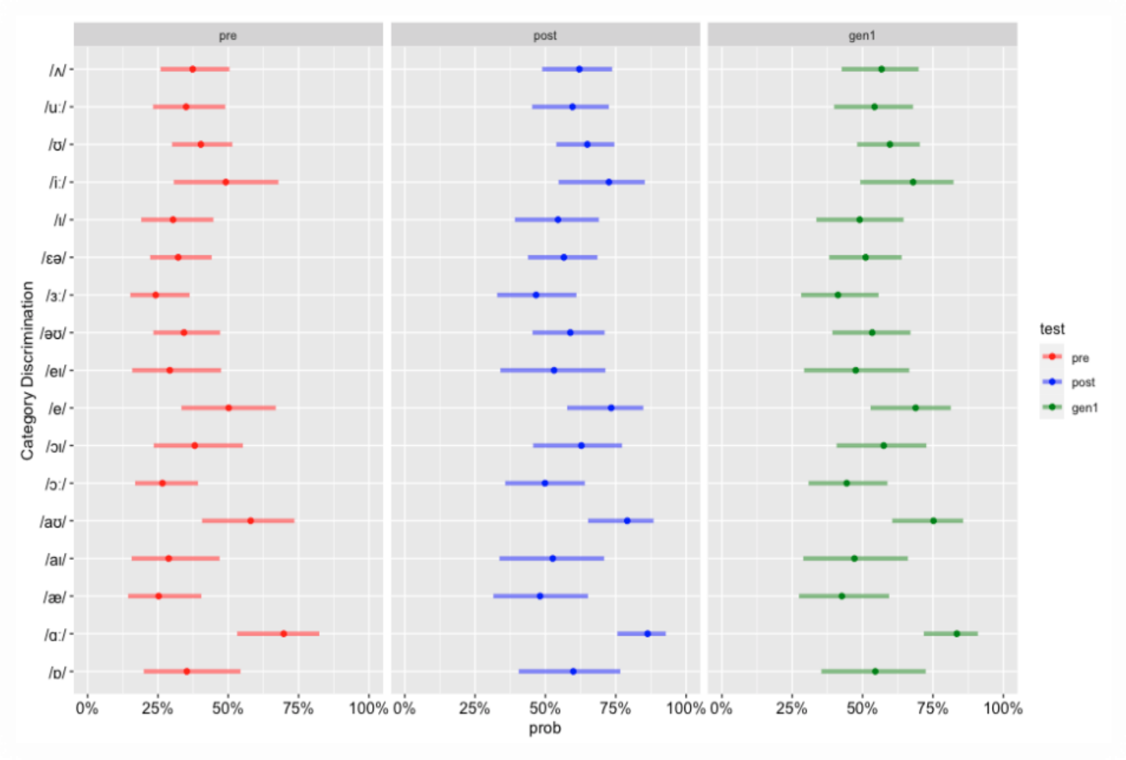


Figure 4.28 The probability of vowels considering the gen2 in relation to pre-and post-test for the CD task (The threshold is 50%).

4.4.4.4 Summary

The below points summarise the findings of the generalisation tests for the CD task:

- The gen1 and gen2 tests of the CD task were significantly more accurate than the pre-test, while the post-test accuracy was comparable. This is consistent with the results of generalisation tests of the AD task.

- During the CD task, groups A, B, and C demonstrated comparable levels of performance on the gen1 and gen2 assessments. This result lines up with the results of the ID and AD tasks. In the context of the three tasks, it can be observed that the inclusion of multiple L1 varieties with a single L2 variety did not provide significant improvements in vowel perception as compared to the use of multiple L1 varieties and a single L1 variety.

- Although most vowels were difficult to distinguish during the pre-test (below the 50% threshold probability), vowel accuracy increased considerably during the two generalisation tests, indicating the training's effectiveness. This is in accordance with the findings of the ID and CD tasks, in which vowel accuracy improved during the generalisation tests compared to the pre-test

4.5. Overview summary of the findings

In summary, the results chapter addresses the research questions presented in the introductory chapter (refer to Section 1.5) by analysing the fixed factors (group, test, and vowel) across three perceptual tasks: Identification (ID), Auditory Discrimination (AD), and Category Discrimination (CD).

Considering the group factor, the performance of three experimental groups was compared across different tests: group A (exposed to a single L1 input), group B (exposed to multiple L1 inputs), and group C (exposed to multiple L1 inputs and a single L2 input). Throughout the pre-, mid-, and post-test phases, groups A, B, and C showed similar performance in ID, AD, and CD tasks. Similarly, the three groups demonstrated comparable performance levels in the generalisation tests (gen1 and gen2) while engaging in the three tasks. While group C demonstrated a modest improvement during the mid-, post-, and generalisation tests compared to groups A and B, this does not reflect better performance, but rather slightly higher starting levels in the pre-test. These results suggest that groups B and C demonstrated the same level of improvement in English vowel perception as group A, indicating successful training across all groups. Thus, it can be concluded that training with diverse accents is not a barrier to learning and is beneficial for learners with low proficiency, proving to be an effective approach in foreign language classrooms.

In regard to the test factor, the chapter investigates the training effects by comparing pre-test, mid-test, and post-test results across groups A, B, and C for different perceptual tasks (ID, AD, and CD). Furthermore, it evaluates the generalisation effects by comparing the generalisation test results (gen1, gen2) with pre-/post-test outcomes.

In ID and CD tasks, groups A, B, and C showed increased probabilities of accurate responses in the post-tests compared to the mid- and pre-tests. Additionally, they demonstrated higher likelihoods of correct responses during the mid-test relative to the pre-test. While performance on the AD task had already reached ceiling levels, slight improvements were still observed across all groups during the mid-test and post-test.

Considering generalisation effects, all groups showed improved accuracy in the gen1 test of the ID task, outperforming the accuracy levels observed in both the pre- and post-tests. In the gen2 test, while the performance exceeded pre-test results, it was consistent with the outcomes observed in the post-test. During the CD task, all groups showed higher accuracy in the two generalisation tests compared to the pre-test, though slightly below the post-test results. Despite hitting ceiling effects (above 90%) in the AD task pre-test, accuracy modestly improved in the two generalisation tests, exceeding pre-test levels and equalling the post-test results. These results indicate that learners successfully transferred their skills to unfamiliar speakers of the same accents (gen1 test) and to speakers of different accents (gen2) across various perceptual tasks, demonstrating effective generalisation of learned abilities.

Furthermore, the chapter analyses vowel performance across the three perceptual tasks (ID, AD, CD). In the ID task, participants experienced moderate difficulty: half of the vowels (/ɔɪ/, /eɪ/, /aʊ/, /æ/, /ɔː/, /ʌ/, /iː/, /aɪ/, /ɑː/) were easily identified with pre-test accuracies above 50% threshold. The remaining vowels (/ɒ/, /e/, /ɪ/, /ɜː/, /ɛə/, /əʊ/, /ʊ/, /uː/) had lower pre-test accuracies, below 50% at the pre-test. In the AD task, all vowels reached a ceiling effect during the pre-test, indicating that learners perceived the task as less challenging and fairly straightforward. In contrast, the CD task was more difficult than the ID and AD tasks, with all vowels except for (/ɑː/, /aʊ/, and /e/) recording below 50% accuracy in the pre-test. Despite varied task difficulties, all vowels improved over time, with greater progress evident in the post-test after 16 training sessions compared to the mid-test after 8 sessions. This highlights the effectiveness of the training and the advantages of extended training periods for Arabic learners with low proficiency.

During the gen1 and gen2 tests, most vowels in the ID task reached accuracy levels above 50%, demonstrating a higher probability of correct responses compared to the pre-test, where nine vowels scored below 50% and eight scored above 50%. Compared to the post-test, vowel accuracy in the gen1 test was higher, while the gen2 test showed consistent levels. In the AD task, vowel discrimination during the two generalisation tests aligned with post-test levels and slightly exceeded those of the pre-test, reflecting the initial ceiling effect. In the CD task, vowels demonstrated improved accuracy in both gen1 and gen2 tests compared to the pre-test, though it remained slightly below the post-test results.

Chapter 5. Discussion

The current chapter review the findings of this thesis and their relation to existing literature. It opens with a summary of the key results, then discusses the impact of incorporating multiple English varieties into the HV phonetics training method. The efficiency of the perceptual tests utilised, namely identification (ID), auditory discrimination (AD), and category discrimination (CD), is then assessed. Following that, generalisation effects are considered. The chapter concludes with an examination of vowel performance before and after training, as well as across generalisation tests. It divides vowels into 'easy' and 'difficult' categories and discusses these distinctions from three perspectives: 1) learner performance, 2) phonological perspective, and 3) phonetic perspective.

5.1 Key findings

To recap, the previous chapter analysed training and generalisation effects within the ID, AD, and CD tasks, concentrating on three factors: groups, tests, and vowels. The key findings for each factor are detailed in the table below.

Identification (ID)		
Groups	**Tests**	**Vowels**
• Groups A, B, and C showed equivalent performance across pre-, mid-, post-, and generalisation tests (gen1, gen2).	• All groups showed increasingly accurate responses in the post-tests relative to the mid- and pre-tests. • In the gen1 test, the three groups showed higher correct response rates than observed in both the pre- and post-tests. In gen2, their performance exceeded pre-test levels and equalled those of the post-test.	• During the pre-test, the vowels (/ɔɪ/, /eɪ/, /aʊ/, /æ/, /ɔː/, /ʌ/, /iː/, /aɪ/, /ɑː/) were easily identified, with accuracies exceeding the 50% threshold, while the vowels (/ɒ/, /e/, /ɪ/, /ɜː/, /ɛə/, /əʊ/, /ʊ/, /uː/) had lower accuracies, falling below 50%. However, all vowels showed improvement over time, with greater gains observed in the post-test compared to the mid-test. • Despite initial classifications of vowels as easy or difficult based on pre-test performance, the majority scored above 50% in

		both gen1 and gen2 tests, demonstrating learners' ability to adapt and retain knowledge. Compared to the post-test results, the gen1 test showed a higher probability of correct responses, while the gen2 test matched them.
Auditory Discrimination		
Groups	**Tests**	**Vowels**
• Groups A, B, and C showed equivalent performance across pre-, mid-, post-, gen1, and gen2.	• Despite reaching peak performance in the AD task, all groups still showed slight improvements during the mid- and post-tests. • Given that the groups' initial pre-test performance peaked, they demonstrated modest improvements in the two generalisation tests, which corresponded with the outcomes of the post-test.	• All vowels reached a ceiling effect during the pre-test, indicating the task was straightforward and easy, yet modest improvements were observed in the mid- and post-tests. • Vowel discrimination in gen1 and gen2 tests aligned with post-test levels and slightly exceeded pre-test results.
Category Discrimination (CD)		
Groups	**Tests**	**Vowels**
• Groups A, B, and C showed equivalent performance across pre-, mid-, post-, gen1, and gen2.	• The three groups showed increasingly accurate responses in the post-tests relative to the mid- and pre-tests. • All groups achieved higher accuracy in the two generalisation tests than in the pre-test, though slightly lower than in the post-test.	• The CD task proved to be the most challenging, with most vowels falling below the 50% accuracy threshold in the pre-test. Despite this, there was a higher probability of correct responses in the post-tests, surpassing those in the mid-test. • In the generalisation tests, all vowels showed enhanced accuracy compared to the pre-test, although they marginally lagged behind the post-test results.

Table 5.1 Summary of Key Findings

5.2 Evaluating the effectiveness of incorporating multiple accents into HV training

The primary goal of this thesis was to determine whether accent variability aids L2 learners' perception of English vowels in the same way that other aspects of variability (multiple speakers and multiple contexts) are beneficial across different perceptual tasks and whether this advantage persists when L1 and L2 varieties are used. The findings revealed that groups A, B, and C showed comparable performance on all three perceptual tasks (ID, CD, and AD), not only during the mid-and post-tests but also during the two generalisation tests (Gen1, Gen2), indicating that exposure to different accents was not problematic for beginner learners. To recap, HV training for group A involved a single L1 variety (SSBE), HV training for group B included a variety of L1 Standard English accents (SSBE, AmE, AusE), while HV training for group C included two L1 varieties (SSBE, AmE) and one L2 variety (Saudi English). Although there was a modest improvement in the performance of group C compared to groups A and B across all assessments, this did not account for higher performance. This is because they also performed better before the training programme (as indicated by the pre-test). The performance of the three groups also remained comparable over the course of the training sessions, with group C showing a slight advantage (see the raw data-based visualisations of the training sessions for the perceptual tasks provided in Figures 5.1, 5.2, and 5.3).

The three figures demonstrate that from sessions 1 to 8, participants were exposed to a specific set of vowels, which was then presented again in sessions 9 to 16. To visually differentiate these vowels across the sessions, unique shapes were assigned to each session. For instance, square shapes are used for sessions 1 and 9 to highlight the same sets of vowels appearing in both. Moreover, session 10 is excluded from the three plots because, in the design for group B, the auditory discrimination task was mistakenly omitted while the category discrimination task was duplicated. As a result, session 2, which shared its vowels with session 10, was also excluded. The visualisation approach implemented in R assigns a unique shape to every two training sessions, which prevents the incorporation of session 2. During the identification task, participants were exposed to the following vowel sets: Sessions 1 and 9 featured /iː ɪ e/ and /e eɪ aɪ/; sessions 3 and 13 included /ɜː ɑː ɔː/ and /ɜː ɛə e/; sessions 4 and 14 presented /uː aʊ əʊ/ and /uː ʊ ʌ/; in sessions 5 and 15, the sets were /aʊ uː ʊ/ and /aʊ ɔɪ əʊ/; and sessions 6 and 16 involved /ɜː eɪ ɛə/ and /ɔː ɑː əʊ/. Additionally, sessions 7 and 8 served as review sessions, revisiting the vowel sets from sessions 1-3 and 4-6, respectively. During the category and auditory discrimination tasks, participants were exposed to the following vowel pairs: In sessions 1 and 9, the pairs included /ɪ e/, /iː e/, /iː ɪ/, /e eɪ/, and /aɪ eɪ/; sessions 3 and 13 featured /ɜː ɑː/, /ɑː ɔː/, /ɜː ɛə/, /ɛə e/, and /ɜː e/; sessions 4 and 14 presented /uː aʊ/, /aʊ əʊ/, /uː ʊ/, /ʊ ʌ/, and /uː ʌ/; sessions 5 and 15 included /aʊ ɔɪ/, /ɔɪ əʊ/, /aʊ uː/, /uː ʊ/, and /aʊ ʊ/; and in sessions 6 and 16, the pairs were /ɜː eɪ/, /eɪ ɛə/, /ɔː ɑː/, /ɑː əʊ/, and /ɔː əʊ/. Sessions 7 and 8 revisited the pairs from sessions 1-3 and 4-6, respectively.

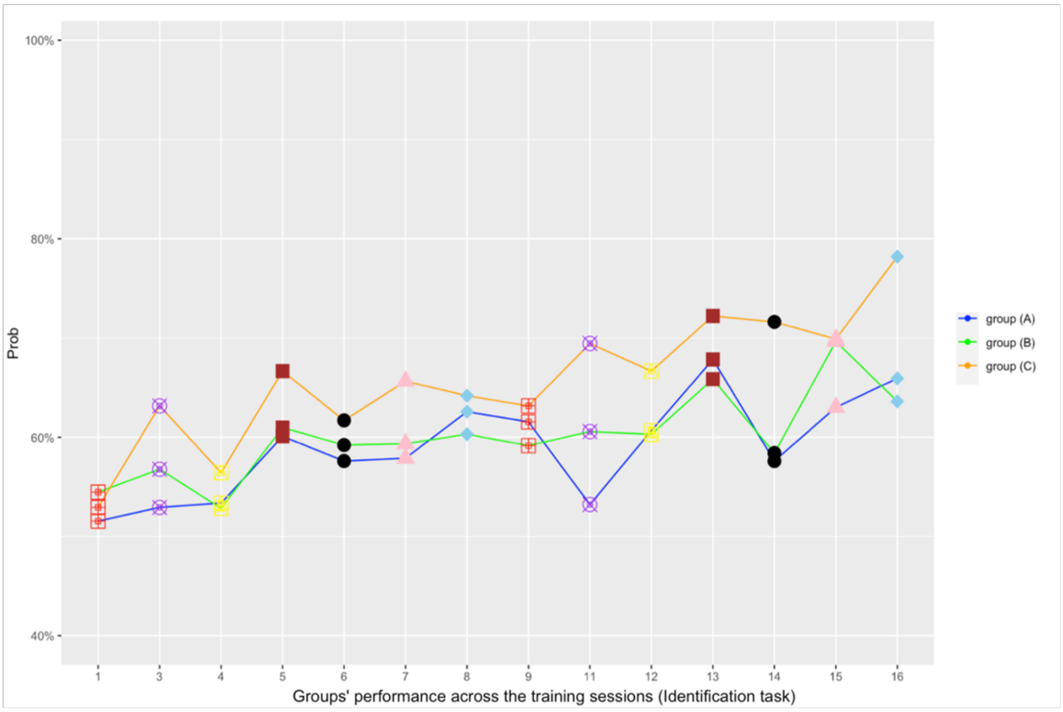


Figure 5.1 Groups' performance during the training sessions for the identification task.

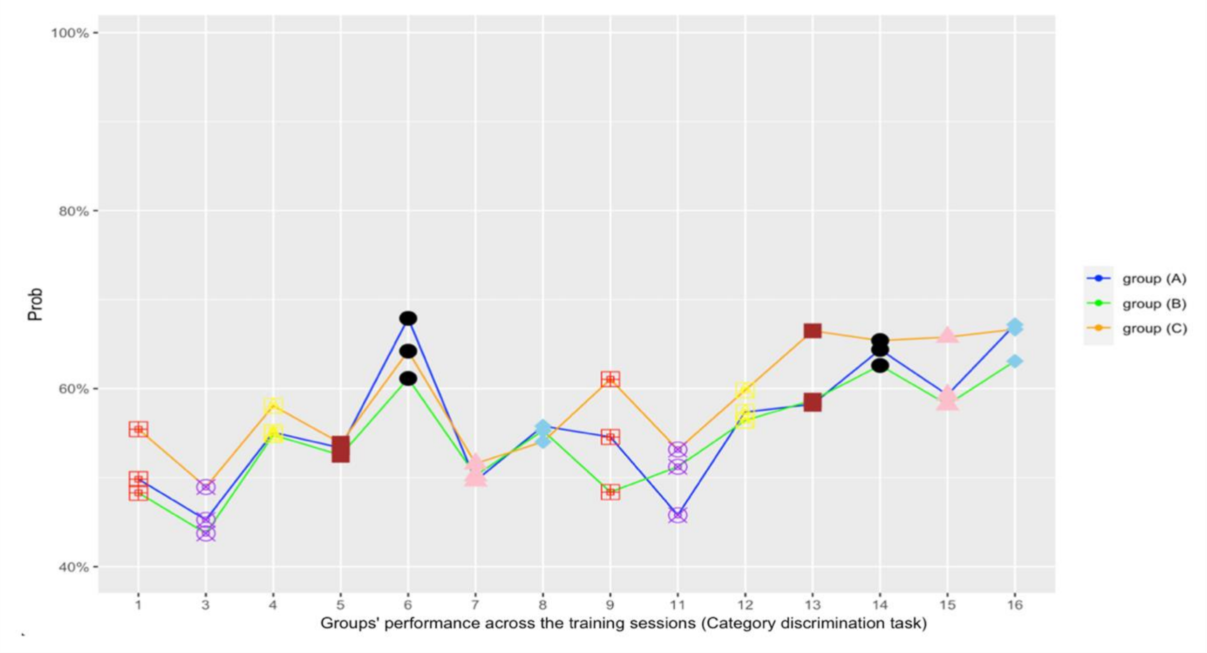


Figure 5.2 Groups' performance during the training sessions for the category discrimination task.

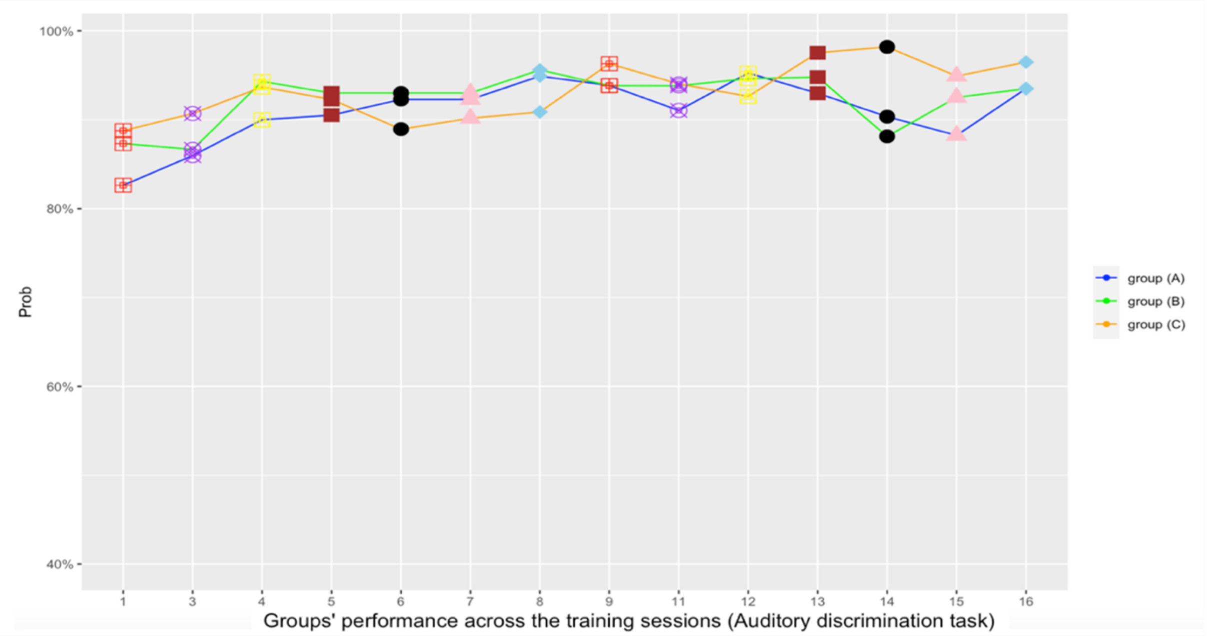


Figure 5.3 Groups' performance during the training sessions for the auditory discrimination task.

The evidence of training success in groups B and C demonstrates that the multiple English varieties used were not at a disadvantage but rather enabled performance that was at the same level as group A. This is remarkable when considering that these were beginner learners with little experience with English. Given the comparable progress across all groups, using L2 and L1 varieties in phonetic training is well-supported, reflecting real-world diversity. Therefore, one could argue that the inclusion of English variation is appropriate for utilisation in FL classroom settings as it does not have a detrimental effect on learners' perceptions. Indeed, these results give validity to the common belief in academic research that L2 learners need to have real-life experiences with English, including exposure to multiple English varieties. This exposure can help prevent potential issues that may arise from only being familiar with one standard L1 variety, such as mistaking regional variations for errors or viewing them as less prestigious forms (Morrow, 2004; Mahboob & Elysa, 2014; Timmis, 2002).

Nonetheless, the equal performance of the three groups was unexpected, given the thesis's hypothesis that exposure to English varieties would lead to better performance than a single variety. Recall that this hypothesis emerged from observations on the advantages of variability in learning, specifically concerning the HV training method, which includes exposure to multiple speakers and diverse contexts (Logan et al., 1991; Iverson et al., 2012). The argument was made that if such variability aids L2 learners, incorporating accent variability could further enhance learning outcomes. A potential explanation for the similar outcomes across the groups in this study could be attributed to the limitation of having only one speaker per accent for groups B and C. In contrast, group A had multiple speakers with the same accent. It appears that participants in groups B and C benefited from the inclusion of multiple accents, even though only one speaker represents each accent. Group A, on the other hand, benefited from the inclusion of multiple speakers despite employing only one accent. In other words, the success of group A, which is similar to that of groups B and C, seems to be influenced by the diversity of speakers. This emphasises the crucial role of speaker variability in the HV training approach. Consequently, if groups B and C were to undergo training involving multiple speakers with different accents, one could reasonably anticipate enhanced outcomes. An aspect that requires further investigation is the extent to which L2 learners would gain from hearing L1 and L2 varieties from multiple speakers compared to hearing various L1 and L2 varieties from a single speaker.

At first, the plan involved having three speakers from each accent in group B (3 SSBE, 3 AmE, 3 AusE) and group C (3 SE, 3 AmE, 3 SSBE), while group A was intended to consist of nine speakers with the same accent (SSBE). The expectation was that incorporating multiple speakers with varying accents would provide learners with a greater advantage compared to various speakers with the same accent. However, due to time constraints and the impacts of the COVID-19 pandemic, the quantity of speakers involved in recording the stimuli was reduced to one per accent. While HVPT may be perceived as more suitable for intermediate or advanced learners, this study demonstrated that even beginner learners could benefit and improve their performance even when variability included multiple accents, speakers and words.

5.3 Evaluating the effectiveness of the training tasks

Another goal of this thesis was to assess the effectiveness of the three perceptual tasks (ID, AD, CD) across different training groups. Overall, the data showed that the training programme successfully addressed all perceptual tasks, notwithstanding the different difficulties associated with each (Refer to the subsections below for a detailed discussion on each task). This observed efficacy can be attributed to various factors that have been highlighted in the literature and that were carefully considered during the training design in an effort to generate a more substantial enhancement in vowel perception. These include the role of variability and immediate feedback in the HV training technique (Logan et al., 1991; Bradlow & Bent, 2003). The ID training task and all discrimination training tasks (i.e., AD, CD, and LingLab vowel game matching game) demonstrated variability: in the training group (A), there is a variety of speakers and words, whereas in groups (B) and (C), this range expands to include accent, speakers, and word variability. Identification and discrimination tasks, according to Shinohara and Iverson (2018), should be effective as long as variability exists.

It is noteworthy that adult beginner learners in this study showed significant improvements despite the inherent variability in the training, a finding that contrasts with previous research, such as Shinohara & Iverson (2021). In their study, although Japanese learners of all age groups—from children to adults—improved their perception of the English /l-r/ contrast through HV training, adolescents showed the most significant gains, particularly in identification, category discrimination, and sensitivity to the F3 formant. This challenges the "younger is better" hypothesis (Flege, 1995), as beginner learners, typically children, did not experience the greatest benefit from variability in training. A key factor that may explain the discrepancy between this study and Shinohara & Iverson (2021) is the cognitive and developmental differences between children and adults. Even as beginners, adults may possess more advanced cognitive strategies and superior attention mechanisms, which enable them to manage variability in training more effectively than younger learners. Additionally, adults are more likely to utilise explicit learning strategies and possess stronger abstract thinking skills, enabling them to recognise and apply patterns in variable phonetic input. This cognitive maturity allows adult learners to benefit more from HV training, while younger learners are still developing these abilities. The SLM-r (Flege & Bohn, 2021) further suggests that learners with more refined and precise L1 phonetic categories are better positioned to detect and form new L2 categories. In this study, adult learners likely had more developed L1 categories, helping them benefit from training despite the variability. Additionally, adult learners, often motivated by academic or professional aspirations, tend to engage more deeply with training, which helps them overcome variability and achieve more consistent improvements than younger learners.

The length of the training is another crucial factor that can significantly improve the performance of the three experimental groups across different perceptual tasks. The training and testing phases were done online over three months[72], with participants completing 16 training sessions and performing pre-, mid-, and post-assessment and generalisation tests. The training period was intentionally lengthened due to the relatively limited vowel inventory in Saudi Arabic, which consists of only eight vowels, as opposed to the denser vowel space in English (Almurashi, 2022). If Arabic learners took part in a few training sessions, the gains in performance across

[72] To the best of my knowledge, this is the longest training study of its kind, with an unprecedented 180 participants. However, the impact of Covid contributed to a significant level of attrition, resulting in a final count of 121 participants. Only 112 of these were examined since being classified as low-proficiency learners.

tasks are expected to be less. For example, Iverson and Evans (2009) observed that while Spanish learners were initially behind German learners in English vowel learning due to their limited vowel system, they were able to catch up after extended training, demonstrating the same progress as German after 15 sessions compared to initial assessments after 5 sessions.

Incorporating a range of perceptual tasks including identification, auditory discrimination, category discrimination and The LingLab vowel matching game offers distinct advantages over the identification-focused HV training used in most studies (e.g., Logan et al., 1991; Lengeris & Hazan, 2010; Iverson et al. (2012). While identification tasks help learners recognise specific sounds, they often do not fully address the ability to distinguish between similar sounds or categorise them accurately. Iverson et al. (2012) found that after eight sessions of HV training focused on identification task, both experienced and inexperienced groups improved in vowel identification accuracy but showed only modest gains in vowel discrimination. By integrating auditory and category discrimination tasks, learners develop a more comprehensive skill set, enhancing their ability to discern fine-grained differences and categorise sounds effectively. This comprehensive approach fosters more robust perceptual skills, essential for accurate language use.

The inclusion of HV production training could have contributed to the observed improvements in perceptual tasks across all groups. Although the comparative effectiveness of HV combined perception and production training versus perception-only training for vowel perception was not tested in this study, a positive impact is expected due to the interconnected aspects of perception and production in the learning process (Flege & Bohn, 2021; Ingram & Park, 1997). While it remains unclear whether the production component specifically facilitated the perceptual gains or simply did not hinder them due to the lack of direct comparisons between training conditions, the consistent improvement from pre-test to mid-test and post-test across all perceptual tasks strongly indicates that production training did not disrupt—and may have enhanced—the development of accurate perceptual representations of English vowels. This finding contrasts with the results of Baese-Berk and Samuel

(2016), who reported that production tasks during training could disrupt perceptual learning. The divergence between these outcomes may stem from differences in task design. In Baese-Berk and Samuel's study, participants were required to produce sounds immediately after each perceptual task within the same trial, likely increasing cognitive load and hindering perceptual consolidation. On the other hand, the current study, similar to those by Alshangiti (2015) and Wong (2014), employed a more structured approach, where perception and production tasks were not intertwined within the same trial, likely reducing cognitive burden and enabling more effective learning.

A key factor that likely contributed to the differing outcomes between the current study and Baese-Berk and Samuel (2016) is the duration and intensity of the training programs. Baese-Berk and Samuel's brief training of 72 ABX discrimination trials over two days may have been insufficient for consolidating perceptual learning, especially with the added cognitive load of production tasks. In contrast, the current study involved a significantly more extensive and varied training tasks. In contrast, the current study involved 1,728 trials over 16 sessions, incorporating a range of incorporating a range of perceptual and production tasks. This more comprehensive approach likely provided richer and more sustained exposure to the target sounds, facilitating more effective perceptual development. It is also possible that participants' production skills improved, although this was not tested in this thesis.

The gradual introduction of vowel sounds across sessions in the current study contributed to more robust and enduring perceptual improvements. This approach contrasts with those of previous studies by Iverson and Evans (2009) and Iverson et al. (2012), which condensed the training into five sessions, each addressing all 14 target vowels. In this study, training was strategically divided into 16 sessions, with each session concentrating on a specific subset of vowels (e.g., /iː/, /ɪ/, /e/; /e/, /eɪ/, /aɪ/ in the first session), ultimately covering 17 target vowels in total. Despite featuring fewer trials per session (108) compared to the 225 trials in the aforementioned studies. The comprehensive design of 1,728 trials distributed over 16 sessions provided consistent and diverse exposure without overwhelming the beginner learners. This structured approach ensured learners received sufficient yet manageable exposure to the target sounds, potentially offering more substantial benefits than the 1,125 trials across five sessions (225 per session) used in previous studies. The following sections discuss the outcomes of the three perceptual tasks and draw connections to the existing body of literature.

5.3.1 Perceptual tasks

This section discusses learners' progress in perceptual tasks, taking both pre-and post-test results into account. While the mid-test scores for all tasks were reported in the results chapter, they are not discussed here. All tasks improved gradually from pre- to mid- to post-tests, but the greatest improvements were seen across pre-and post-tests, so the section focuses on these findings.

5.3.1.1 Identification

The identification task was moderately difficult for SA learners, with their pre-training performance ranging from 47% to 52%. The data showed a significant improvement of 11–12% from the pre- to post-tests, which is a great achievement considering the training was delivered online to low-proficiency participants. The observed improvement here aligns with 10–15% progress often reported in research studies that solely utilise HV identification training methods (Lengeris & Hazan, 2010; Iverson et al., 2005; Bradlow et al., 1999; Logan & Pruitt, 1995; Lively et al., 1993; Jamieson & Morosan, 1986). Furthermore, the findings are consistent with Shinohara and Iverson (2018), who included both identification and discrimination tasks in their HV training and reported a 12% improvement in the identification of /r/-/l/ English sounds. This implies that incorporating DIS and ID tasks into HV training, as done in the current thesis and Shinohara and Iverson (2018), does not provide a greater increase in phoneme identification compared to previous research that focused only on ID tasks (e.g., Iverson et al., 2005; Bradlow et al., 1999; Lively et al.,1993).

The training design was guided by a review of the literature, revealing that training is not entirely task specific. For instance, Carlet and Cebrian (2019) tested two

HV training—one incorporating an ID task and the other an AX discrimination task—both of which improved vowel identification, though the discrimination task demonstrated a comparatively smaller improvement than the identification task. The inherent complexities of the identification task's orthography could explain the modest progress seen in vowel identification in the current study despite the use of various perceptual and production tasks. This is especially relevant considering that the participants in this thesis have a low level of proficiency, which could potentially impede their ability to handle orthographic intricacies. Undeniably, there exists a lack of research studies on the examination of the role of orthography and word frequency in identification tasks.

However, if the entire training phase is dedicated to the identification task, learners can be predicted to achieve higher identification results than those documented in this thesis as long as the number of training sessions meets learners' needs. For instance, research conducted by Iverson and Evans (2009) and Iverson et al. (2012), all of which focused on identification tasks throughout HV training, revealed vowel identification improvement that was greater than those stated in this thesis. In these studies, learners from various backgrounds experienced notable increases in identification accuracy: Spanish learners experienced a 20% increase after 15 training sessions, according to Iverson & Evans (2009), while inexperienced French learners improved by 25% after 8 sessions and their experienced peers by 17% in Iverson et al. (2012). Consequently, allocating training sessions exclusively to identification activities, rather than splitting the session time between different types of tasks (identification, production, and discrimination), may allow learners to correlate auditory stimuli with their accurate representations more adeptly. This, in turn, could potentially alleviate the difficulties associated with orthography. However, it is logical to anticipate discrepancies in L2 sound identification as a result of methodological variances, including differences in the number of training sessions and the development of a robust training programme tailored to the requirements of the learners. Moreover, variations in the outcomes of sound identification across different research studies could potentially be attributed to factors such as the motivation of the learners and their proficiency in the L2.

5.3.1.2 Discrimination tasks

The AD task was found to be quite simple for SA learners, with their performance before training hitting a ceiling effect, varying between 95% and 97%. The performance of groups A, B, and C demonstrated a slight improvement in their post-test scores, with a 3% to 4% increase compared to their initial pre-test scores. By way of explanation, learners had not been sufficiently challenged by the AD task, given they had encountered a ceiling effect before the start of the training, leaving little room for significant improvement in their auditory discrimination abilities. To reduce the ceiling effect, it can be advantageous to add noise (e.g., multi-talker babble) to the stimuli. This addition would heighten the task's difficulty, requiring learners to exert greater cognitive effort. Indeed, prior research has used noise as a criterion to test the robustness of training methods, verifying if they are adequate for the intricate and fluctuating situations that one commonly encounters outside of the controlled training environment (e.g., Lengeris & Hazan, 2010; Cutler et al., 2004; Iverson & Evans, 2007b).

For the CD task, participants in the three experimental groups initially faced great difficulties, as evidenced by pre-test scores ranging from 34% to 42% below the 50% threshold. Despite the increased memory demand and greater uncertainty in the stimuli of the CD discrimination task (Strange & Dittmann, 1984), which required learners to compare three distinct stimuli, the training effectively enhanced the learners' competence in this task. Impressively, the performance levels of groups showed a noteworthy improvement of 25% from the pre- to the post-tests, with scores ranging from 59% to 67%. The improvements observed in the CD task were considerably higher, reaching double the improvements seen in the ID task. This discrepancy may arise from the participants' lower language proficiency, which could potentially lead to confusion when it comes to the orthography in the ID task. They listen to a sound and are then required to select from a set of three written words, a process that can be complicated by the intricacies of spelling. Conversely, the CD task focuses solely on sensory discrimination, requesting participants to pick out a uniquely vowel-sounded word from three, free of any spelling-related distractions.

A plausible explanation for the observed improvement in the CD could be ascribed to the incorporation of the *LingLab vowel-matching game*, which exhibits a greater resemblance to the CD task compared to the ID task. Learners actively participated in this game seven times following the conclusion of each training session. Such involvement is likely to improve their vowel discrimination skills more than their identification skills. However, in their study, Iverson et al. (2023) used a gamified approach to improve the perception of SSBE vowels in Japanese adult learners, providing a contrasting perspective. They contended that, even though the memory card game was more aligned with the CD task than the ID task, their findings showed that Japanese adult learners performed much better in the identification task than the category discrimination test. This indicates that the efficacy of game-based learning methods in improving particular perceptual capabilities is unclear, highlighting the need for additional research into how different elements of these games contribute to various aspects of language learning.

The findings obtained from the CD task present a notable divergence from the conclusions drawn in several training studies (e.g., Heeren & Schouten, 2008; Iverson & Evans, 2009; Lengeris & Hazan, 2010; Alshangiti, 2015) that found HV training improved vowel identification but did not necessarily improve learners' ability to discriminate among vowels. The observed differences in results can be ascribed to methodological differences: the referenced studies exclusively incorporated ID tasks in their training procedures, whereas the training developed for this thesis included not just identification but also discrimination and production tasks. In addition, the significant improvement seen in the CD task in this study challenges previous views on the limited effectiveness of auditory discrimination tasks (Strange & Dittmann, 1984). Rather, the findings align with Shinohara and Iverson (2018), who showed that including both ID and DIS tasks improves discrimination of L2 consonant contrasts, such as the English /r/-/l/, as long as variability exists (Shinohara & Iverson, 2018). Although this thesis shows significant progress in the CD task and even exceeds the improvements reported in Shinohara and Iverson's 2018 research, it is unclear whether the addition of HV discrimination training is the primary contributor to learners' improved English vowel discrimination. It is possible that a combination of factors, including training duration, variability, immediate feedback, the inclusion of discrimination, identification, and production training, all contributed to this improvement.

5.4 Generalisation effects

The findings of two generalisation tests, Gen1 and Gen2[73], are discussed in this section. When the results of the generalisation tests match or exceed the post-test scores and differ from the pre-test results, generalisation is deemed to have occurred, following Carlet and Cebrian (2019). Recall that Gen1 involves new words produced by new talkers of familiar accents (SSBE, SE), whereas Gen2 features these same new words but is produced using unfamiliar accents (Indian, Chinese). The tasks in the generalisation tests were the same as those in the pre/mid/post-tests (i.e., ID, AD, CD).

5.4.1 Generalisation test I

The results showed that learners' performance across the three training groups (A, B, C) was equally capable of successfully transferring their training to new words produced by new talkers of familiar English varieties (SSBE, SE). Specifically, the groups demonstrated a 20% to 21% improvement in the ID and CD tasks and a 2% to 4% improvement in the AD task[74] compared to their performance in the pre-test. When compared to the post-test findings, their performance increased considerably in the ID, with a significant 9% to 10% improvement. The accuracy remained similar to the post-test results in the AD task while decreased by only 3% in the CD task. Generalisation appears feasible in the CD task, as the results show a slight deviation from the post-test scores and a significant difference from the pre-test results. In comparison to previous studies, the Gen1 findings demonstrated greater

[73] The goal was to perform generalisation evaluations following the post-test. However, the online mode of the training resulted in irregular test timings, with participants completing the tests at different intervals (for example, some after 3 days from the post-test, others after 7 days, and even after two weeks). Despite this, upon initial inclusion of the time variable in the analysis model, it was found to have an insignificant effect on the test results.

[74] The lower generalisation performance observed in the AD task is not due to difficulty, but rather, as previously stated, it reflects the learners' performance having reached a ceiling effect.

generalisation effects. For example, Lively et al. (1994) found that training with talker-specific information significantly aided in identifying new words spoken by a familiar talker more than new words spoken by an unfamiliar talker. Accent-specific information in Gen1 (SSBE heard by all three groups, SA presented only to group C but potentially recognisable to groups A and C as the speaker shared the SA accent with the learners) was effective in assisting learners in identifying and discriminating new words spoken by unfamiliar speakers.

Furthermore, the generalisation effects observed in the current study exceed those reported by Carlet & Cebrian (2019) and Shinohara & Iverson (2018). They analysed the difference between pre-test and post-test outcomes as a measure of generalisation to real words. Carlet & Cebrian found that the ID group improved their vowel identification by 15%, while the AX-DIS group improved by 1.5%. On the other hand, Shinohara and Iverson (2018) combined ID and DIS training and found a 12-15% improvement in the perception of /r/-/l/ sounds during the ID and CD tasks. Though the Gen1 test of the current study was taken at different times by learners following the post-test, the gains in learning were 20% to 21% for the ID and CD tasks across the three groups compared to the pre-test results. Again, the greater improvements observed in the ID and CD tasks for Gen1 in comparison with previous studies could be attributed to the factors (described in section 5.2) that were carefully considered during the training design phase.

5.4.2 Generalisation test 2

Like the Gen1 test, the three groups in the Gen2 test were equally capable of applying their training to new words spoken by speakers of different varieties of English (Indian, Chinese). Vowel identification scores increased by 12% during Gen2 compared to pre-test results but remained comparable to post-test scores. Though significant, this progress in the ID task of Gen2 is less than the improvement observed in Gen1, where there was a 20% to 21% increase over pre-test scores and a 9% to 10% rise compared to post-test scores. This suggests that learners were better at identifying vowels produced by new speakers of familiar English varieties than unfamiliar English varieties. Even though the words in Gen2 were the same as those in Gen1, it is possible that participants needed to put forth more mental effort to adapt to unfamiliar accents while also focusing on accurately matching the heard stimuli with the correct written word. Despite hearing unfamiliar accents in the Gen2 ID task, learners' improvement (12%) is not significantly different from the 15% generalisation gain observed in previous studies (such as Carlet & Cebrian, 2019; Shinohara & Iverson, 2018), in which learners were exposed to familiar accents and the difference between pre-test and post-test results was used as a measure of generalisation.

The results of the AD task in Gen2 showed a slight distinction from the pre-test scores (3% to 4%), yet they remained in line with the post-test scores. This pattern closely aligns with the results of the Gen1 test, suggesting that the modest gains indicate a limited potential for further improvement among learners who have already reached a ceiling effect. Similarly, progress in the CD task during Gen2 corresponds to that of Gen1: there was a 21 to 22% improvement in Gen2 compared to pre-test scores, while the performance was marginally lower (3%) than the post-test results. When performing the CD task of the Gen2 test, students have to process unfamiliar accents and identify odd vowels among three different words. Despite these demands, the CD task improves more than the ID task, especially when compared to pre-test scores. One might expect generalisation to new English varieties to be lower for the CD task than for the ID task, given that initial CD performance was below the 50% threshold, with scores ranging from 34% to 42%. This contrasts with the higher pre-test ID task scores, which ranged from 47% to 52%. The significant discrepancy between the CD task results in Gen2 and the pre-test indicates its suitability for learners with limited proficiency, as it eliminates any potential spelling confusion. Although the results of the three tasks (ID, CD, AD) did not outperform the post-test results, generalisation emerges since none showed a significant drop: ID and AD performance were consistent across both tests, and the CD task in Gen2 showed only a tiny decline when compared to the post-test. The tasks used in different evaluations (pre-, post-, and generalisation tests) were discussed in this section; the following section addresses vowel performance across these tests.

5.5 Assessing vowel sounds

Data showed that vowel perception improved consistently across tasks (ID, AD, and CD). Notably, the post-test, administered after 16 training sessions, revealed greater improvement than the mid-test, administered after 8 sessions. In assessing the AD task, all vowels were readily identified with scores above 90% in the pre-test, indicating a ceiling effect and perceived ease by learners. Despite this high initial performance, vowel discrimination improved even more in the post-test and generalisation tests. Given the high level of proficiency demonstrated by vowels in the AD task, a detailed discussion is not included. The following sections focus on vowels that are easy or difficult in the ID and CD tasks only, classifying them based on learner performance. The findings are then discussed from a phonetic and phonological standpoint.

5.5.1 Easy Vs Difficult vowels based on learners' performance

5.5.1.1 Vowel Identification task

In the vowel identification task, Arabic participants easily identified the vowels /ɔɪ/, /eɪ/, /aʊ/, /æ/, /ɔː/, /ʌ/, /iː/, /aɪ/, and /ɑː/, achieving success rates above 50% in the pre-test. On the other hand, the vowels /ɒ/, /e/, /ɪ/, /ɜː/, /ɛə/, /əʊ/, /ʊ/, and /uː/ were more challenging in the initial assessment, with their perception rates being below 50%. These findings demonstrate that phonetic similarities and differences between Arabic and English constitute significant factors in how well Arabic learners perceive English vowel sounds. Except for the /ʌ/ and /aɪ/, the performance of Arabic learners in this study is consistent with Alshangiti's (2015) findings of SA low-proficiency learners. The table below contrasts the vowels categorised as easy and difficult in both the current study and Alshangiti's (2015) research. Both studies target low-proficiency SA learners[75], who identify the speech of SSBE speakers. However, it is reasonable to expect differences in individual learner performance. Table 5.2 shows differences in the accuracy of easy and difficult vowel identification across studies, but the overall accuracy of vowel classification should be reliable.

[75] Alshangiti (2015) also involved SA learners with high proficiency; however, since the current research exclusively focuses on low-proficiency learners, only the low-proficiency participants from Alshangiti are being considered.

	Easy Vowels	Accuracy		Easy Vowels	Accuracy
Current study	/ɔɪ/	69%	Alshangiti (2015)	/ɔɪ/	56%
	/eɪ/	69%		/eɪ/	59%
	/aʊ/	68%		/aʊ/	59%
	/æ/	66%		/æ/	79%
	/ɔː/	65%		/ɔː/	62%
	/iː/	58%		/iː/	74%
	/ɑː/	51%		/ɑː/	85%
	/ʌ/	59%			
	/aɪ/	55%			
	Difficult Vowels	Accuracy		Difficult Vowels	Accuracy
Current study	/uː/	44%	Alshangiti (2015)	/uː/	36%
	/əʊ/	39%		/əʊ/	18%
	/ʊ/	38%		/ʊ/	51%[76]
	/ɛə/	38%		/ɛə/	18%
	/ɜː/	38%		/ɜː/	44%
	/e/	36%		/e/	69%[77]
	/ɪ/	30%		/ɪ/	8%
	/ɒ/	14%		/ɒ/	3%
				/ʌ/	31%
				/aɪ/	46%

Table 5.2 Comparison of easy and difficult vowel identification in the current study and Alshangiti (2015): High accuracy in easy vowels is highlighted in green, while low accuracy in difficult vowels is highlighted in red.

Furthermore, the current study's classification of vowels into easy and difficult categories corresponds to Algethami's (2023) findings on production data. Algethami used acoustic analysis to examine the production of 12 SSBE monophthongs (ɪ iː ʊ uː ʌ æ ɑː ɒ ɔː e ɜː ə) by upper intermediate SA learners in the UK, focusing on accuracy in comparison to native SSBE speakers. Even though the emphasis was on accuracy rather than intelligibility, this study sheds light on which vowels are easier or more difficult for learners to produce (See section 2.7). The vowels /ʊ uː ɪ e ɜː ɒ/ were difficult to produce. By way of explanation, Arabic learners exhibited equivalent spectral

[76] Despite the identification rate of the vowel /ʊ/ being 51%, it was considered a challenging vowel since the vowels /uː/ and /əʊ/ were confused with /ʊ/ in 54% and 36% of cases, respectively.

[77] Despite the vowel /e/ having an identification rate of 69%, it was considered a difficult vowel due to the frequent confusion with /ɪ/, which occurred in 72% of cases.

patterns in their production of English vowels /ʊ/ and /uː/. Their productions of /e/ and /ɪ/ were quite similar. Algethami also found that the /ɜː/ vowel had distinct spectral characteristics, with a more forward and elevated position than English speakers. Additionally, its articulation was closer to /e/. There is a slight overlap between the pronunciation of /ɒ/ and Arabic /oː/. On the other hand, the vowels /iː ɔː ɑː ʌ aː/ were easy to produce by Arabic speakers. Learners' production of the vowel sound /iː/ aligned with Arabic and English productions of the same vowels. The production of the /ɔː/ vowel was quite straightforward, as it resembles the Arabic /oː/ sound. There were no major spectral differences in the production of /ʌ/ between learners and English speakers. In addition, learners often pronounced English vowels /æ/, /ɑː/, and /ʌ/ like their native vowel /aː/.

The categorisation of vowels into 'easy' and 'difficult' categories, derived from the current study's vowel identification outcomes and those reported by Alshangiti (2015), as well as the production data from Algethami (2023), support each other in the conclusion that phonetic similarities and differences significantly influence Arabic learners' ability to perceive and produce English vowel sounds. This emphasises the necessity of discussing shared and unshared vowels between Arabic and English from a phonological and phonetic standpoint (See section 4.5.2).

Importantly, the current study's training effectively improved the identification of vowels, regardless of their difficulty level. For example, vowels like /ɔɪ/ and /eɪ/, which were identified with a high accuracy of 69% in the pre-test, saw a 10% increase in the post-test. Similarly, the more challenging vowels, such as /ɒ/ and /e/, with initial pre-test accuracy rates of 14% and 30%, showed improvements of 7% and 12% respectively, in the post-test. The training's effectiveness is further validated by the fact that participants were consistently able to effectively retain and apply the knowledge acquired during the training in the generalisation tests. Vowel identification accuracy in the Gen1 test outperformed both the pre-test and post-test results. The increase in vowel identification was approximately twice as significant when comparing the Gen1 test results to the pre-test than the post-test. In the gen2 test, the accuracy of all vowels improved compared to the pre-test, while the results were like those of the post-test.

One of the studies conducted by Alshangiti (2015) aimed at improving the perception and production of 14 SSBE vowels among SA learners, utilising three distinct training approaches: HVPT, hybrid training (a mix of perception and production), and production training, which included ID and CD tasks. However, a direct comparison between her study and the current research is difficult due to the varied methodologies used in each. The present study treated vowels, tasks, and groups as fixed effects, whereas Alshangiti focused on variables including time, proficiency, and type of training. While Alshangiti reported overall improvements in vowel identification for the hybrid and HVPT training groups, with less improvement in the production group, the study did not detail the accuracy for each vowel in the identification task. This section discusses vowels based on learners' performance on the ID task; the following section addresses vowel performance on the CD.

5.5.1.2 Category discrimination task

Categorising vowels as 'difficult' or 'easy' in the context of the CD task is less applicable. This is because most vowels[78] were found difficult to discriminate by learners during the pre-test phase, with their accuracy falling below the 50% threshold. The vowels /ɑː/, /aʊ/, and /e/ were the only ones to yield a high accuracy rate, with /ɑː/ reaching a score of 70%, and /aʊ/ and /e/ each scoring 50%. Despite this, there was a significant improvement in vowel discrimination during the post-test, ranging from 16% to 25%. Moreover, in the generalisation assessments, vowel accuracy outperformed pre-test results, albeit just slightly less than post-test performance. Based on the findings, it is evident that learners faced challenges with the task initially, as indicated by their low performance in the pre-test. However, they showed consistent improvement in vowel discrimination as they advanced through the training. The improved vowel accuracy seen in both the post-test and generalisation tests demonstrates positive progress in their discrimination skills.

[78] To mitigate issues of convergence and high collinearity that arise when including vowel pairs, individual vowels are represented as fixed effects in the study. Meanwhile, vowel pairs are incorporated as random effects.

The more perceptual ability to distinguish phonetic contrasts in the post-test and generalisation test seems to be primarily affected by the HV training method used in this study, which exceeds the results reported by Iverson (2012) and Alshangiti (2015). Iverson et al. focused on French learners of varying proficiency levels, while Alshangiti explored Saudi Arabic learners with diverse proficiencies. Despite using the same stimuli, neither study demonstrated significant improvements in vowel discrimination from the pre-test to the post-test[79]. Importantly, the two studies used the same discrimination task, in which participants were shown three stimuli pronounced by three different speakers. Two stimuli were identical, while the other was distinct; the learners' task was to identify a different word. On the other hand, the CD task used in the current study presented an even greater challenge. Participants were asked to identify the odd vowel sound from three different words spoken by the same speaker. The more challenging nature of the discrimination task in the current study, compared to those by Iverson and Alshangiti, further underscores the effectiveness of the training method employed in this research.

The improved accuracy in identifying and distinguishing vowels, as observed in this study, is in line with L2 theoretical frameworks (e.g., SLM (Flege, 1995), SLM-r (Flege & Bohn, 2021), and L2LP (Escudero, 2005)) that acknowledge perceptual learning as a continuous process throughout an individual's life. These models also emphasise the importance of L2 experience in shaping learners' perception and production of L2 sounds. The observed improvements indicate that learners' progress is strongly linked to their engagement in extensive training and exposure to the target language input. In addition, the significant enhancement in vowel accuracy across various tasks highlights the effectiveness of the training approach used in this study, indicating that as long as the training is well-designed, improved performance is achievable.

[79] /i/–/ɪ/, /eɪ/–/ɪ/, /eɪ/–/aɪ/, /eɪ/–/e/, /ɑ/–/a/, /ɑ/–/ɒ/, /ɒ/–/ʌ/, /ɜ/–/ʌ/, /ɜ/–/əʊ/, /ɜ/–/ɔ/, /ɑ/–/ɔ/, /ɒ/–/ɔ/, /ɔ/–/aʊ/, /əʊ/–/ɔ/, /əʊ/–/ɒ/, /əʊ/–/aʊ/, /u/– /aʊ/.

5.5.2 Shared Vs unshared vowels based on phonological and phonetic perspectives

This section categorises the shared and unshared vowels (i.e., those in English but not in Arabic) based on phonological and phonetic differences between the two languages. This approach offers a detailed framework that incorporates abstract linguistic patterns with the concrete physical properties of speech sounds. Vowel classifications derived from phonological and phonetic perspectives are compared exclusively to learners' performance in the ID task. The comparison excludes their performance on the AD and CD tasks. This is because, in the pre-test of the AD task, all vowels were easily identified, resulting in ceiling effects, whereas in the CD task, the majority of vowels were difficult to perceive at the pre-test level, as evidenced by accuracy rates below 50%.

Arabic learners' perception of English vowels is discussed within the context of the original SLM (Flege, 1995) and its revised version, SLM-r (Flege & Bohn, 2021) because they are well-suited to evaluating the perception of individual sounds rather than sound contrasts. Their emphasis on position-sensitive allophones, guided by the principle that listeners match incoming sounds to pre-existing long-term memory representations, ensures efficient processing during real-time speech. Although other models like PAM (1995) and L2LP (2005) are influential in understanding L2 perception, they focus primarily on L2 contrasts rather than the perception of individual sounds. Since the analysis in this thesis treats individual vowels as a fixed factor, the SLM and SLM-r are the most appropriate theoretical frameworks for explaining the vowel perception abilities of EFL students.

SLM and SLM-r propose that learners instinctively and automatically link L2 sounds with their corresponding L1 phonetic categories. These models suggest that the greater the phonetic dissimilarity between an L2 sound and its closest L1 equivalent, the higher the likelihood of forming a new phonetic category for the L2 sound. While the models primarily emphasise phonetic learning, they also implicitly account for phonological similarities and differences in mapping L2 sounds onto L1 categories and forming new phonetic categories. Given this framework, it might be expected that Arabic learners would find it challenging to perceive L2 vowels that are phonetically and phonologically similar to their L1 vowels, while potentially finding it easier to perceive entirely new L2 vowels. The underlying assumption here is that the greater the dissimilarity between Arabic and English vowels, the easier the perception should be. However, this assumption requires careful consideration for several reasons.

First, classifying vowels as easy or difficult to perceive based solely on phonetic and phonemic dissimilarity is not entirely practical, as both the SLM and SLM-r suggest that learners do not necessarily form new phonetic categories for all L2 sounds. Some L2 sounds may closely resemble L1 sounds to the point that substituting one for the other might go unnoticed by monolingual speakers of the target language. For instance, the English vowel /i:/ is relatively easy for Arabic learners to perceive, despite its phonetic and phonological similarity to the Arabic /i:/. Additionally, the formation of new phonetic categories is influenced by factors beyond mere dissimilarity between L1 and L2 sounds. SLM-r provides a more nuanced explanation by considering additional factors including the quality and quantity of L2 input, the accuracy of L1 categories at the time of first exposure to the L2, intersubject variability, and endogenous factors like auditory acuity, early-stage auditory processing, and auditory working memory.

Rather than predicting which vowels are phonetically and phonological easy or difficult to perceive, a more effective approach is to identify and compare the shared and unshared vowels with learners' performance. Table 5.3 and 5.4 below present a summary of vowels categorised as shared and unshared from both phonological and phonetic perspectives based on findings from studies on Saudi Arabic and English (Almurashi et al., 2023; Algethami, 2023; Alshangiti & Evans, 2015), alongside the performance of Arabic learners in the current thesis. This approach provides deeper insights into how these similarities and differences influence vowel perception in L2 acquisition. By examining patterns in learners' successes and challenges, a more precise understanding of the specific difficulties faced by Arabic learners in mastering English vowels can be developed. Additionally, it informs the creation of more targeted teaching strategies by highlighting which vowels may need intensified instructional focus, thereby enhancing overall language learning outcomes.

Phonological viewpoint	**L2 vowels**	**Compared to Arabic learners' performance**
Shared vowels	/iː/	Easy to perceive
	/ɪ/	Difficult to perceive
	/uː/	Difficult to perceive
	/ʊ/	Difficult to perceive
	/ɑː/	Easy to perceive
	/æ/	Easy to perceive
Unshared vowels	/ʌ/	Easy to perceive
	/e/	Difficult to perceive
	/ɒ/	Difficult to perceive
	/ɔː/	Easy to perceive
	/ɜː/	Difficult to perceive
	/eə/	Difficult to perceive
	/əʊ/	Difficult to perceive
	/aɪ/	Easy to perceive
	/aʊ/	Easy to perceive
	/eɪ/	Easy to perceive
	/ɔɪ/	Easy to perceive

Table 5.3 Comparison of phonologically shared and unshared vowels with learners' performance

Phonetic viewpoint	**L2 vowels**	**Compared to Arabic learners' performance**
Shared vowels	/iː/	Easy to perceive
	/ɑː/	Easy to perceive
	/æ/	Easy to perceive

	/ʌ/	Easy to perceive
	/ɔː/	Easy to perceive
	/aɪ/	Easy to perceive
	/aʊ/	Easy to perceive
	/eɪ/	Easy to perceive
	/ɔɪ/	Easy to perceive
Unshared vowels	/ɪ/	Difficult to perceive
	/uː/	Difficult to perceive
	/ʊ/	Difficult to perceive
	/e/	Difficult to perceive
	/ɒ/	Difficult to perceive
	/ɜː/	Difficult to perceive
	/eə/	Difficult to perceive
	/əʊ/	Difficult to perceive

Table 5.4 Comparison of phonetically shared and unshared vowels with learners' performance

Tables 5.3 and 5.4 demonstrate that the categorisation of vowels as either simple or challenging for learners is closely tied to their phonetic similarities and differences rather than their phonological properties. Specifically, L2 vowels that are phonetically shared, /iː/, /ɑː/, /æ/, /ʌ/, /ɔː/, /aɪ/, /aʊ/, /eɪ/, and /ɔɪ/ are more easily perceived by learners. These vowels have phonetic correspondences with Arabic vowels: The vowels /iː/ and /aɪ/ are similar to the Arabic /iː/, the vowel /ɑː/ is akin to Arabic /aː/, the vowel /æ/ matches Arabic /a/, the vowel /ʌ/ is close to Arabic /a/ or /aː/, the vowels /ɔː/ and /ɔɪ/ mirror Arabic /oː/, the vowel /aʊ/ is comparable to Arabic /uː/, while the vowel /eɪ/ resembles Arabic /eː/. On the other hand, phonetically distinct vowels, including /ɪ/, /uː/, /ʊ/, /e/, /ɒ/, /ɜː/, /eə/, and /əʊ/, pose more challenges for learners. Being at the initial stages of language learning, low-proficiency Arabic learners may have not fully developed the ability to represent sounds at an abstract phonological level. Predictions derived from the findings of this thesis indicate that English vowels phonetically similar to Arabic ones tend to be perceived more easily by learners. In contrast, English vowels that do not share phonetic traits with learners' native vowels appear to be more challenging for learners to perceive. The training, which includes three groups exposed to different English varieties and consistently yields similar results, reinforces the conclusion that learners in each group more easily acquire vowels that are phonetically similar to those in their native language.

Despite categorising vowels as easy or difficult based on learners' performance and recognising the challenges posed by phonetic dissimilarities, significant improvement was observed across all vowel categories over time. This trend aligns with the SLM's claim that L2 learners gradually discern phonetic differences between their L1 and L2 as they gain more exposure and experience. The SLM-r further elaborates on this by highlighting the critical role of input quality and quantity in shaping L2 phonetic categories, suggesting that learners' progress is influenced not just by the dissimilarity between L1 and L2 sounds, but also by the nature of the input they receive. In this study, although the FL learners had limited natural interactions outside of classroom settings, the structured and consistent training they received was sufficient to enhance their perception of English vowels. This emphasises the vital role of well-designed language instruction in FL contexts and reinforces the importance of both the quality and quantity of exposure in L2 acquisition within educational settings.

The original SLM focused on whether L2 learners could attain native-like proficiency in L2 phonetic categories, comparable to that of L1 speakers. However, the SLM-r shifts away from this expectation, acknowledging that perfect mastery is unlikely due to the interaction between learners’ L1 and L2 phonetic systems and the distinct phonetic input they receive. Consequently, this study encourages learners to aim for target-like perception and production, aligning more closely with the SLM-r's emphasis, rather than the native-like standards highlighted in the SLM. While promoting the pursuit of target-like proficiency, assessing learners' progress towards this goal is challenging without detailed speech analysis. To achieve a target-like level of perception and production of L2 sounds, learners still need to develop new categories for sounds absent in their native tongue (to prevent misunderstanding) and adjust existing ones similar to their first language. While it is uncertain if learners have completely achieved target-like sound categories, the noticeable improvements in identification and discrimination tasks seen in the post-test and the generalisation tests suggest that learners are indeed progressing in the right direction. Additionally, these improvements indicate that the training has been beneficial in helping listeners cope with diverse stimuli, in line with the conclusions of Iverson et al. (2005) and Iverson & Evans (2009). Therefore, the HV training method is valuable for assisting learners in engaging in everyday conversations, where they will likely encounter a wide range of sounds and expressions. The next section addresses the limitation of this thesis and puts forward recommendations for future work.

5.6 Limitations and future work

Building on the findings of the current thesis, the following limitations and recommendations are proposed for future research:

- **Development of teaching materials:**

  In light of the findings from the current study, it is recommended to enrich teaching materials with a diverse range of English varieties as spoken by both L1 and L2 speakers. This approach would greatly enhance the language learning curriculum, making it more inclusive and effective. By doing so, Learners would be better able to adapt to the various phonetic idiosyncrasies prevalent across different English-speaking regions, enhancing their overall language proficiency and adaptability.

- **The inclusion of multiple speakers from different varieties of English:**

  More research is needed to determine how incorporating various accents into the HV training method affects the perception and production of English vowels. This investigation is required to validate and expand on the findings of the current study. One potential limitation of this study is that it only includes one speaker when presenting L1 and L2 English varieties. While group A had the opportunity to listen to various speakers of the same English variety, groups B and C were exposed to one speaker presenting three different varieties.

Considering the comparable outcomes obtained by the three groups, it would be interesting for future research to include speakers with various accents. This can be accomplished by contrasting multiple speakers with the same English accent versus various speakers with different English accents.

- **Analysis of learner production:**

  The current study has shed light on learners' perceptual abilities by examining identification and discrimination tasks. To further promote understanding, future research should also consider learners' speech output. This would necessitate research into the relationship between perception improvements and progress in spoken language production, particularly in terms of attaining target-like performance.

- **Development of target-like proficiency measurement criteria:**

  Future research should focus on the development and validation of reliable measurements to assess target-like perception and production. This could involve developing assessment tools that more accurately align with the practical objectives of L2 learners, particularly those not aiming for native-like fluency but rather functional and intelligible communication.

- **Exploring phonetic and phonological characteristics**:

  Given that learners' performance on easy and difficult vowels tends to align closely with whether the vowels are phonetically shared or unshared between Arabic and English, future research could benefit from a more in-depth examination of the phonetic and phonological features of learners’ L1 and L2. Such research could explore how these characteristics influence not only vowel perception but also the broader aspects of L2 acquisition, potentially identifying patterns that could inform more targeted and effective language teaching strategies. This expanded focus could also shed light on how specific phonetic and phonological attributes either facilitate or hinder the learning process for learners with different linguistic backgrounds.

- **Introducing background noise in testing stimuli**

  Integrating background noise, like multi-talker babble, into the testing stimuli provides a more authentic representation of everyday language environments. This method prepares students to handle practical situations in which English is used in crowded or demanding environments, such as crowded areas or less-than-ideal communication channels. Such an environment poses a beneficial challenge to learners, enhancing their auditory processing skills and leading to improved concentration and comprehension, even in less-than-ideal listening situations. Furthermore, adding noise to test environments can help mitigate the ceiling effect observed in this thesis' auditory discrimination task. By including this element of authenticity, learners are less likely to achieve perfect scores solely based on ideal testing conditions, necessitating more cognitive effort. This allows for a more accurate assessment of individuals' true language comprehension skills in practical situations.

- **Implementation of the retention test**

  The current thesis successfully recruited some participants from each training group to assess their long-term memory retention. Specifically, there were 16 participants from group A, 17 from group B, and 18 from group C. The intention was to evaluate their retention abilities four to five weeks after they completed the training. However, there was a significant difference in the timing of when learners took the test. Some participants took the test a month after finishing the training, while others took it three, four, or even five months later. This significant variation in test timing raised methodological concerns. Despite the sufficient number of participants in each group for a valid comparison, the inconsistency in the timing of the retention test could potentially impact the accuracy and reliability of measuring real long-term retention. Consequently, it was decided not to include learners' retention test performance. Indeed, the variability in test scheduling should be viewed as a limitation of administering the test online. Since this training study was carried out during the COVID-19 period, it was not possible to manage the training or testing phases on campus. Future research should include retention assessments of learners' long-term retention of perceptual skills and, potentially, production abilities. A more structured testing schedule is recommended to improve the reliability of these assessments. Conducting these tests on campus rather than online, if possible, could help reduce the significant time differences observed in the current study. This controlled environment would almost certainly result in more consistent and accurate assessments of long-term learning outcomes.

- **Conducting HV training on campus as part of a phonetic course**

  Using the HV training method in a phonetic course on campus can provide a dependable and controlled learning environment. The current study's online format had advantages, including the ability to recruit more participants and provide learners with more freedom and flexibility, but it also had some drawbacks. A significant issue was that, even though participants recorded their voices for the production task of the training, some of these recordings were not successfully uploaded to the Labvanced platform. This issue was frequently caused by the various devices used by participants, resulting in inconsistencies in the quality of submissions and technical compatibility. As a result, these learners had to terminate the training. However, in an on-campus setting, a consistent technological setup can be provided, ensuring that all students have access to the same quality of equipment and resources. This consistency aids in the avoidance of technical issues that plague online submissions. It would be impossible to have all learners perform simultaneously on campus for the production task. Instead, instructors should schedule individual time slots for each participant to complete the task independently.

- **Improving the LingLab vowel matching game**

  As previously mentioned, the time constraints imposed by the RSE led to specific limitations in the Linglab vowel matching game. A notable constraint is the lack of functionality to monitor users' progression, thereby requiring them to create an account and log in for any progress monitoring. Additionally, users have unrestricted access to all sets (set1, set2, set3), which may not align with the intended training protocol. Future research endeavors aimed at improving the effectiveness of the game should duly recognise and take into account these aforementioned limitations. This includes creating a more robust progress-tracking system and potentially restricting access to specific sets based on the user's training stage or group, resulting in a more structured and focused learning experience.

- **Designing perceptual and production tasks from scratch**
  While platforms such as Labvanced are user-friendly and enable the creation of various studies without coding, they typically fall short in providing the necessary customisation and flexibility for intricate tasks and feedback. For example, a major limitation in the production task is that the software only captures final responses, without recording the preceding steps. This limitation makes it difficult to verify whether participants strictly followed the instructions (record first, listen to a model's production, then record again). To address these limitations in future research, developing production and perceptual tasks from scratch using coding skills would allow for significantly greater flexibility and innovation in task design and feedback mechanisms. Learning to code could change the method of researchers in the field of L2 learning, allowing them to create a wide range of tasks tailored to even the most unique or specific requirements of their studies. This approach has the potential to significantly advance current research methodologies in the field of L2 learning research beyond the constraints of existing platforms.

- **Developing an application to improve FL learning techniques.**
  Future research should consider bringing together all the tasks employed in this study into a user-friendly application. This approach would optimise access to the training for learners, guaranteeing seamless navigation and boosted engagement. It also presents a substantial enhancement compared to the existing system, wherein training/ testing sessions are accessed via links and subjected to the variability of different devices. By developing a cohesive

application, researchers can reduce technical obstacles and optimise the overall efficacy of the training procedure, mainly when conducting it remotely.

The following chapter presents a succinct overview of this thesis, summarising its objectives and outcomes.

Chapter 6. Conclusion

This thesis provides empirical evidence supporting the efficacy of the HV training method in improving Arabic FL learners' perception of English vowels. The application has been shown to benefit FL learners whose L1 is dominant and who have fewer opportunities to interact with L1 and L2 English speakers outside of the classroom. The successful outcomes achieved not only showcase the effectiveness of the method in enhancing English vowel perception but also validate its potential in boosting EFL teaching through the provision of materials grounded in solid theoretical foundations. Despite the extensive time and effort required in the training's design, including the recruitment of speakers for recording stimuli and the creation of tasks/activities, the investment is justified by the notable improvement in the perceptual skills of a large number of students, including their capacity to process different varieties of English. This type of training is adaptable, as it can be delivered on campus with adequate computer access and/or online as independent homework assignments for learners.

The primary purpose of this thesis is to assess the success of the HV method, particularly in its incorporation of multiple L1 and L2 English varieties. To date, much research on HV training has concentrated on training materials delivered in a single L1 English variety, typically SSBE or AmE. In the present study, all three training groups experienced equivalent levels of improvement: "A" received only SSBE, "B" was exposed to a range of L1 Standard English accents (including SSBE, AmE, and AusE), and "C" was exposed to two L1 varieties (SSBE, AmE) and one L2 variety (Saudi English). These findings strongly advocate for including L2 and L1 varieties in phonetics instruction, mirroring the diversity found in the real world and promoting social justice by accurately representing the role of different varieties in the classroom. Although this study did not incorporate a systematic qualitative evaluation, some participants from each of the three training groups (A, B, C) initiated communication with the researcher one month after the training concluded. They reported that their understanding of movies and news improved, and they were able to process English vowels more efficiently than before participating in the training; this potentially demonstrated the training's practical efficacy in enhancing their perceptual abilities.

When building the HV training framework, previous research focused primarily on a single perceptual task (identification). The procedure resulted in significant improvements in the learners' ability to identify vowels. However, it became clear that their ability to discriminate vowels was not significantly improved. Recognising this limitation, this thesis's vital aim was to thoroughly evaluate the efficacy of combining identification, auditory discrimination and category discrimination tasks across three training groups. This exploration showed that the training regimen tackled all the designated perceptual tasks despite each posing its own set of difficulties. In addition, the findings indicated that participants in each of the three training groups (A, B, C) consistently demonstrated the ability to apply their acquired knowledge to new words spoken in both familiar (Saudi, SSBE) and unfamiliar English varieties (Indian, Chinese). However, their ability to generalise their knowledge to familiar varieties was greater than to unfamiliar varieties. Therefore, the training outcomes suggest that to see improvement on both fronts, it is essential to incorporate stimulus variability along with integrating identification and discrimination tasks.

Additionally, the study assessed vowel performance across different assessments and tasks. Significant improvement in vowel accuracy for both identification and discrimination tasks was evident in the post-test, following 16 training sessions, in comparison to the mid-test results achieved after just 8 sessions. This implies that the duration of the training is vital for enhancing the ability of Arabic learners to perceive English vowels. The importance of this need is highlighted by the fact that Arabic has a fewer number of vowels than English. As a result, Arabic learners require additional practice to adjust to the more intricate vowel system in English. The *auditory discrimination task* showed a high level of vowel accuracy during the pre-test, implying a ceiling effect. However, despite this excellent initial performance, there was even more improvement in vowel discrimination in both the post-test and generalisation tests. The majority of vowels in the *category discrimination task* did not exceed the 50% accuracy threshold. Yet, there was a noticeable improvement in vowel discrimination during the post-test and generalisation tests. Regarding the vowel identification task, 9 vowels (/ɔɪ/, /eɪ/, /aʊ/, /æ/, /ɔː/, /ʌ/, /iː/, /aɪ/, /ɑː/) were determined to be easily perceivable, with an accuracy rate exceeding the 50% threshold in the pre-test. On the other hand, 7 vowels (/ɒ/, /e/, /ɪ/, /ɜː/, /ɛə/, /əʊ/, /ʊ/, /uː/) proved to be more difficult to discern, with accuracy falling below the 50% threshold. Interestingly, there was notable progress in identifying both the easy and the difficult vowels. Further, participants successfully utilised the knowledge they gained during training in the generalisation tests.

The findings of this thesis underscore the significant influence of phonetic similarities between Arabic learners' native language and English on vowel perception and acquisition. The consistent performance across the three groups of Arabic learners revealed that their ability to perceive and acquire vowels is more closely aligned with the degree of phonetic similarity between the vowels in English and their L1, rather than with broader phonological properties. Specifically, vowels that are phonetically shared between Arabic and English were more easily perceived and acquired, while those that lacked such phonetic similarities posed greater challenges. These results lead to predictions that phonetic resemblance plays a crucial role in the initial stages of vowel learning, guiding learners' ability to process and internalise new sounds in a second language. These insights contribute to a deeper understanding of the factors that facilitate or hinder L2 vowel acquisition and highlight the importance of considering phonetic factors in the development of effective language teaching strategies.

Importantly, the current thesis points out the value of L2 learners striving for target-like perception and production rather than exclusively focusing on achieving native-like proficiency. Current L2 models predominantly evaluate learners' perception and production based on a comparison with native/L1 speakers, primarily using acoustic analysis or asking L1 speakers to judge learners' productions. Shifting this traditional focus to establish more achievable goals focused on intelligibility and target-like performance can benefit learners due to its practicality. Assessing target-like performance can be done by performing a vowel identification task which evaluates learners' production rather than perception[81]. For example, listeners identify learners' production of the word 'hid' from a closed set of options, each paired with two common words: 'heed' (sheet, teeth), 'head' (pen, desk), and 'hid' (kid, rich). While this task mirrors those in previous studies (e.g., Iverson et al., 2012) which focused on auditory intelligibility, the significant difference here is in the selection of judges. Prior research typically assessed learners' productions for intelligibility to L1 listeners (often SSBE), but this study advocates for including both L1 and L2 listeners to ensure that learners' productions are intelligible to a broader range of listeners. However, the measurement of target-like performance lacks a solid foundation and has not been thoroughly investigated in existing research. Thus, it is vital to conduct comprehensive studies on how to accurately assess and evaluate learners' performance to go beyond the traditional native-like benchmarks. Given that this study is centered on enhancing learners' perceptual skills, assessing their progress towards achieving target-like proficiency is difficult without a thorough analysis of their spoken language. Yet, the significant progress seen in identification and discrimination tasks during the post-test and generalisation tests is supportive.

[81] The key distinction between using vowel identification to assess perception and production lies in the sample. The former focuses on L2 learners who select the correct answer to a spoken stimulus, while the latter involves English listeners who match learners' productions to a close set of stimuli.

References

Abdoh, E. M. A. (2011). A Study of the Phonological Structure and Representation of First Words in Arabic. *PhD Thesis, University of Leicester, Leicester, UK*.

Abou Haidar, L. (1994). Norme linguistique et variabilité dialectale : analyse formantique du système vocalique de l'arabe standard. *Revue de Phonétique Appliquée*, *110*, 1–22.

Agresti, A. (2012). *Categorical data analysis*. John Wiley & Sons.

Algethami, G. (2023). The production of English monophthong vowels by Saudi L2 speakers. *Poznan Studies in Contemporary Linguistics*, *59*(3), 475–492. https://doi.org/10.1515/psicl-2022-1073

Alghamdi, M. M. (1998). A spectrographic analysis of Arabic vowels: A cross-dialect study. *Journal of King Saud University*, *10*(1), 3–24.

Al-khresheh, M. H. (2020). The Impact of Cultural Knowledge on Listening Comprehension of EFL Learners. *English Language Teaching*, *2*(3), 349–371.

Almbark, R. (2012). *The Perception and Production of SSBE vowels by Syrian Arabic learners: The Foreign Language Model*. PhD thesis, University of York.

Almbark, R., & Hellmuth, S. (2015). *Acoustic analysis of the Syrian vowel system*. 1–5.

Almegren, A. (2018). Saudi Students' Attitude towards World Englishes. *International Journal of Applied Linguistics and English Literature*, *7*(4), 238.

Almurashi, W. (2022). *Acoustic cues in production of English vowels by Hijazi Arabic learners - Newcastle University*. PhD Thesis, University of Newcastle.

Almurashi, W., Al-Tamimi, J., & Khattab, G. (2019). *Static and dynamic cues in vowel production in Hijazi Arabic*. *147*(4), 3468–3472.

Almurashi, W., Al-Tamimi, J., & Khattab, G. (2020). Static and dynamic cues in vowel production in Hijazi Arabic. *The Journal of the Acoustical Society of America*, *147*(4), 2917–2927.

Almurashi, W., Al-tamimi, J., & Khattab, G. (2023). *Static and dynamic features of English monophthongal vowels by Hijazi Arabic L2 learners*.

Al-Nasser, A. S. (2015). Problems of English Language Acquisition in Saudi Arabia: An Exploratory-cum-remedial Study. *Theory and Practice in Language Studies*, *5*(8), 1612.

Alotaibi A. N. (2018). *The role of native language dialect on the perception of L2 English vowels*. PhD Thesis, Indiana University.

Alseadan, M. S. (2021). Improving Listening Skills of Saudi EFL Learners. *International Journal of Social Science and Human Research*, *04*(12), 3682–3687.

Al-Seghayer, K. M. (2023). *Status and Functions of English in Saudi Arabia*.

Al-Shaibani, A. (2023). Challenges in Teaching Pronunciation to Saudi Female Learners at Taif University. In *Arab World English Journal*. M.A. Thesis, Taif University.

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

Al-Tamimi, J. (2007). *Static and dynamic cues in vowel production: A cross dialectal study in Jordanian and Moroccan Arabic.*

Al-Tamimi, J. (2022). *Session 5: Advanced statistical analyses*. https://jalalal-tamimi.github.io/R-Training/Session5-AnalysingData_advanced.nb.html#5_Cumulative_Logit_Link_Models

Alzamil, J. (2021). Listening Skills: Important but Difficult to Learn. *Arab World English Journal*, *12*. https://doi.org/10.2139/SSRN.3952957

Ani, S. H. (1978). The development and distribution of the Qaaf in Iraq. Readings in Arabic linguistics. *Indiana University Linguistics Club*, 103–112.

Aoyama, K. (2003). Perception of syllable-initial and syllable-final nasals in English by Korean and Japanese speaker. *Second Language Research*, *19*(3), 251–265.

Authority General Entertainment. (2023). *الهيئة العامة للترفيه[General Entertainment Authority]*. https://www.gea.gov.sa/en/about-us/

Baese-Berk, M. M., Bradlow, A. R., & Wright, B. A. (2013). Accent-independent adaptation to foreign accented speech. *The Journal of the Acoustical Society of America*, *133*(3), EL174–EL180.

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.

Barriuso, T. A., & Hayes-Harb, R. (2018). High Variability Phonetic Training as a Bridge from Research to Practice. *CATESOL Journal*, *30*(1), 177–194.

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting Linear Mixed-Effects Models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Bench, J., & Bamford, J. (1979). *Speech-Hearing Tests and the Spoken Language ofHearing- Impaired Children*. Academic Press, London.

Bent, T., Bradlow, A. R., & Smith, B. (2007). Phonemic errors in different word positions and their effects on intelligibility of non-native speech. In *Language experience in second language* (pp. 331–347).

Bernard, J. R. (1970). Toward the acoustic specification of Australian English. *STUF - Language Typology and Universals*, *23*(1–6), 113–128. https://doi.org/10.1524/STUF.1970.23.16.113

Best, C. T. (1993). Emergence of Language-Specific Constraints in Perception of Non-Native Speech: A Window on Early Phonological Development. *Developmental Neurocognition: Speech and Face Processing in the First Year of Life*, 289–304.

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

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. *John Benjamins*, 13–34.

Boersma, P., & Weenink, D. (2021). *Praat: doing phonetics by computer [Computer program]. Version 6.1.51*. https://www.fon.hum.uva.nl/praat/

Bohn, O. S., & Flege, J. E. (1992). The Production of New and Similar Vowels by Adult German Learners of English. *Studies in Second Language Acquisition*, *14*(2), 131–158.

Bohn, O. S., & Munro, M. J. (2007). *Language Experience in Second Language Speech Learning: In honor of James*. John Benjamins Publishing.

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.

Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., & Tohkura, Y. (1999). Training Japanese listeners to identify English /r/and /1/: Long-term retention of learning in perception and production. *Perception and Psychophysics*, *61*(5), 977–985. https://doi.org/10.3758/BF03206911

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*(2), 707–729. https://doi.org/10.1016/J.COGNITION.2007.04.005

Brekelmans, G. (2020). Phonetic vowel training for child second language learners: the role of input variability and training task. In *Doctoral thesis, UCL (University College London).*

Browman, C. P., & Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology*, *6*(2), 201–251.

Bruthiaux, P. (2003). Squaring the circles: issues in modeling English worldwide. *International Journal of Applied Linguistics*, *13*(2), 159–178.

Buck, G. (2001). *Assessing Listening*. Cambridge University Press.

Burnham D., Kim J., Davis C., Ciocca V., Schoknecht C., & Luksaneeyanawin S. (2011). Are tones phones? *Journal of Experimental Child Psychology*, *108*, 693–612.

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

Carlet, A. Cebrian, J. (2019). Assessing the Effect of Perceptual Training on L2 Vowel Identifi cation, Generalization and Long-term Effects. *A Sound Approach to Language Matters*, 91–119.

Carney, A. E., Widin, G. P., & Viemeister, N. F. (1977). Noncategorical perception of stop consonants differing in VOT. *The Journal of the Acoustical Society of America*, *62*(4), 961–970.

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

Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. London: Harper & Row.

Clark, J., Yallop, C., & Fletcher, J. (2007). *An Introduction to Phonetics and Phonology*. Blackwell Publishing.

Cook, V. (1999). Going beyond the Native Speaker in Language Teaching. *TESOL Quarterly*, *33*(2), 185–209. https://doi.org/10.2307/3587717

Council of Europe. (2018). *Common European Framework of Reference For Languages: Learning, Teaching, Assessment. Companion Volumew With New Descriptors*.

Cox, F. (2006). The Acoustic Characteristics of /hVd/ Vowels in the Speech of some Australian Teenagers. *Australian Journal of Linguistics*, *26*(2), 147–179.

Cox, F., & Fletcher, J. (2017). *Australian English pronunciation and transcription*. Cambridge University Press.

Cox, F. M. (1996). *An Acoustic Study of Vowel Variation in Australian English*. Ph.D. dissertation, Macquarie University.

Cox, F., Palethorpe, S., Tabain, M., Fletcher, J., Grayden, D., Hajek, J., & Butcher, A. (2010). *Broadness variation in Australian English speaking females*. *34*(1), 175–178.

Crystal, D. (1992). *Introducing linguistics*. Penguin English.

Crystal, D. (2003). *English as a Global Language*. Cambridge university press.

Derwing, T. M. & Munro, M. J. (2015). *Pronunciation Fundamentals: Evidence-Based Perspectives for L2 Teaching and Research*. John Benjamins Publishing Company.

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

Eisenstein, M., & Berkowitz, D. (1981). The effect of phonological variation on adult learner comprehension. *Studies in Second Language Acquisition*, *4*(1), 75–80.

Elvin, J., Williams, D., & Escudero, P. (2016). Dynamic acoustic properties of monophthongs and diphthongs in Western Sydney Australian English. *The Journal of the Acoustical Society of America*, *140*(1), 576–581.

Escudero, P. (2005). *Linguistic perception and second-language acquisition: Explaining the attainment of optimal phonological categorization*.

Escudero, P. (2009). Linguistic perception of “similar” L2 sounds. Phonology in perception. In *Phonology in perception* (pp. 152–190).

Escudero, P., & Boersma, P. (2002). *The subset problem in L2 perceptual development: multiple- category assimilation by Dutch learners of Spanish*. 208–219.

Escudero, P., & Boersma, P. (2004). Bridging the gap between L2 speech perception research and phonological theory. *Studies in Second Language Acquisition*, *26*(4), 551–585.

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–EL260.

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

Fairbanks, G. (1960). *Voice and articulation drillbook* (2nd ed.). New York: Harper & Row.

Field, J. (2005). Intelligibility and the Listener: The Role of Lexical Stress. *TESOL Quarterly*, *39*(3), 399–423. https://doi.org/10.2307/3588487

Flege, J. E. (1987). The production of “new” and “similar” phones in a foreign language: Evidence for the effect of equivalence classification. *Journal of Phonetics*, *15*(1), 47–65.

Flege, J. E. (1995). *Second language speech learning: Theory, findings and problems*. 233–277.

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

Flege, J. E., & MacKay, I. R. (2010). *"Age" effects on second language acquisition*. 113–118.

Flege, J. E., Munro, M. J., & MacKay, I. R. A. (1995). Factors affecting strength of perceived foreign accent in a second language. *The Journal of the Acoustical Society of America*, *97*(5), 3125–3134.

Flege, J. E., & Port, R. (1981). *Cross-Language Phonetic Interference: Arabic to English*. *24*(2), 125–146.

Foote, J. A., Holtby, A. K., & Derwing, T. M. (2011). Articles Survey of the Teaching of Pronunciation in Adult ESL Programs in Canada. *TESL Canada Journal*, *29*(1), 1–22.

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.

Fuhrmeister, P., & Myers, E. B. (2017). Non-native phonetic learning is destabilized by exposure to phonological variability before and after training. *The Journal of the Acoustical Society of America*, *142*(5), EL448–EL454.

Galloway, N., & Rose, H. (2015). *Introducing Global Englishes*. Routledge.

Gerrits, E. (2001). *The categorisation of speech sounds by adults and children: a study of the categorical perception hypothesis and the development weighting of acoustic speech cues*.

Grenon, I., Kubota, M., & Sheppard, C. (2019). The creation of a new vowel category by adult learners after adaptive phonetic training. *Journal of Phonetics*, *72*, 17–34.

Hahn, L. D. (2004). Primary Stress and Intelligibility: Research to Motivate the Teaching of Suprasegmentals. *TESOL Quarterly*, *38*(2), 201. https://doi.org/10.2307/3588378

Harmer, J. (2001). The practice of English language teaching. In 3rd (Ed.), *Jeremy. Harmer*. Pearson Education.

Harrington, J., & Cassidy, S. (1994). Dynamic and Target Theories of Vowel Classification: Evidence from Monophthongs and Diphthongs in Australian English. *Language and Speech*, *37*(4), 357–373.

Harrington, J., Cox, F., & Evans, Z. (1997). An acoustic phonetic study of broad, general, and cultivated Australian English vowels. *Australian Journal of Linguistics*, *17*(2), 155–184.

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.

Hetzron, R. (1997). *The Semitic Languages*. Taylor & Francis.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America*, *97*(5), 3099–3111.

Hillenbrand, J. M. (2013). Static and dynamic approaches to vowel perception. *Vowel Inherent Spectral Change*, 9–30.

Hillenbrand, J. M., & Nearey, T. M. (1999). Identification of resynthesized /hVd/ utterances: Effects of formant contour. *The Journal of the Acoustical Society of America*, *105*(6), 3509–3523.

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

Honorof, D. N., McCullough, J., & Somerville, B. (2000). *Comma gets a cure: A diagnostic passage for accent study*. https://www.dialectsarchive.com/comma-gets-a-cure

Huensch, A. (2019). Pronunciation in foreign language classrooms: Instructors' training, classroom practices, and beliefs. *Language Teaching Research*, *23*(6), 745–764.

Huthaily, K. (2003). *Contrastive phonological analysis of Arabic and English*.

Ingram, J. C. L., & Park, S. G. (1997). Cross-language vowel perception and production by Japanese and Korean learners of English. *Journal of Phonetics*, *25*(3), 343–370.

International Phonetic Association. (1999). *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press.

Ioup, G., Boustagi, E., El Tigi, M., & Moselle, M. (1994). Reexamining the critical period hypothesis: A case study of successful adult SLA in a naturalistic environment. *Studies in Second Language Acquisition*, *16*, 73–98.

Isaacs, T., & Trofimovich, P. (2012). Deconstructing comprehensibility. *Studies in Second Language Acquisition*, *34*(3), 475–505.

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

Iverson, P., & Evans, B. G. (2009). Learning English vowels with different first-language vowel systems II: Auditory training for native Spanish and German speakers. *The Journal of the Acoustical Society of America*, *126*(2), 866–877.

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., 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.

Iverson, P., Herrero, B. P., & Katashima, A. (2023). Memory-card vowel training for child and adult second-language learners: A first report. *JASA Express Letters*, *3*(1), 15202.

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., & Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition*, *87*(1), B47–B57. https://doi.org/10.1016/S0010-0277(02)00198-1

Jenkins, J. (2003). *World Englishes: A Resource Book for Students*. Psychology Press.

Jenkins, J. (2009). English as a lingua franca: interpretations and attitudes. *World Englishes*, *28*(2), 200–207.

Jongman, A., Herd, W., & Al-Masri, M. (2007). *Acoustic correlates of emphasis in Arabic*. *16*, 913–916.

Jurafsky, D., & Martin, J. (2009). *Speech and Language Processing* (2nd ed). Prentice Hall.

Kachru, B. B. (1986). The power and politics of English. *World Englishes*, *5*(2–3), 121–140. https://doi.org/10.1111/J.1467-971X.1986.TB00720.X

Kachru, B. B. (1991). Teaching world Englishes. In *English across cultures* (Vol. 2, Issue 2).

Kachru, B. B. (1992). World Englishes: approaches, issues and resources. *Language Teaching*, *25*(1), 1–14.

Kachru, Y., & Smith, L. E. (2008). *Cultures, Contexts, and World Englishes*. Routledge.

Kang, O. (2010). ESL learners' attitudes toward pronunciation instruction and varieties of English. *Pronunciation in Second Language Learning and Teaching*, *1*(1), 105–118.

Kartushina, N., & Martin, C. D. (2019). Talker and Acoustic Variability in Learning to Produce Nonnative Sounds: Evidence from Articulatory Training. *Language Learning*, *69*(1), 71–105.

Kent, R. D., & Read, C. (2002). *The acoustic analysis of speech*. Singular Thomson Learning.

Khattab, G. (2007). *A phonetic study of gemination in Lebanese Arabic*. 153–158.

Khattab, G., & Al-Tamimi, J. (2008). Durational cues for gemination in Lebanese Arabic. *Language and Linguistics*, *11*(22), 39–56.

Kim, Y. H., & Hazan, V. (2010). *Individual variability in the perceptual learning of L2 speech sounds and its cognitive correlates*.

Kirkpatrick, A., Deterding, D., & Wong, J. (2008). The international intelligibility of Hong Kong English. *World Englishes*, *27*(3–4), 359–377.

Kirkpatrick, Andy. (2007). World Englishes: Implications for international communication and English language teaching. In *Cambridge University Press.* Cambridge University Press.

Kopczynski, A., & Meliani, R. (1993). The vowels of Arabic and English. *Papers and Studies in Contrastive Linguistics*, *27*, 183–192.

Kramsch, C. (1997). Guest Column: The Privilege of the Nonnative Speaker. *PMLA*, *112*(3), 359–369. https://doi.org/10.1632/S0030812900060673

Kuhl, P. K. (1994). Learning and representation in speech and language. *Current Opinion in Neurobiology*, *4*(6), 812–822.

Kuhl, P. K. (2000). A new view of language acquisition. *National Academy of Sciences*, *97*(22), 11850–11857.

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. https://doi.org/10.1098/RSTB.2007.2154

Kuhl, P. K., & Iverson, P. (1995). Linguistic experience and the “perceptual magnet effect.” In *Speech perception and linguistic experience: Issues in cross-language research* (pp. 121–154).

Ladefoged, P., & Disner, S. F. (2012). *Vowels and Consonants*. Willey Blackwell.

Ladefoged, P., & Johnson, K. (2015). *A Course in Phonetics* (7th ed). Wadsworth.

Ladefoged, P., & Maddieson, I. (1996). The sounds of the world’s languages. In *Oxford: Blackwell*. Oxford: Blackwell.

Lammertyn, J., Leuven, K. U., Nicolaï, B., De Ketelaere, B., & Molenberghs, G. (2003). Generalised linear mixed model for multicategorical responses: the effect of UV-C treatment on strawberry sepal quality. *Acta Horticulturae*, 495–504.

Lehiste, I., & Peterson, G. (1961). Transitions, Glides, and Diphthongs. *The Journal of the Acoustical Society of America*, *33*(3), 268–277.

Lengeris, A. (2009). Individual differences in second-language vowel learning. *Doctoral Thesis, UCL (University College London).*

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

Lengeris, A., & Hazan, V. (2007). *Cross-language perceptual assimilation and discrimination of southern British English vowels by Greek and Japanese learners of English*. 1641–1644.

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), 3757–3768.

Lenneberg, E. H. (1967). *Biological foundations of language*. Wiley-Blackwell.

Levis, J. M. (2005). Changing Contexts and Shifting Paradigms in Pronunciation Teaching. *TESOL Quarterly*, *39*(3), 369. https://doi.org/10.2307/3588485

Levis, J. M. (2018). *Intelligibility, oral communication, and the teaching of pronunciation*. Cambridge University Press.

Levis, J. M., Sonsaat, S., Link, S., & Barriuso, T. A. (2016). Native and Nonnative Teachers of L2 Pronunciation: Effects on Learner Performance. *TESOL Quarterly*, *50*(4), 894–931. https://doi.org/10.1002/tesq.272

Levis. J, Moyer. A. (2014). *Social Dynamics in Second Language Accent*. De Gruyter Mouton.

Lively, S. E., Logan, J. S., & Pison, D. B. (1993). Elastic Constants of Isotropic Cylinders using Resonant Ultrasound. *Journal of the Acoustical Society of America*, *94*(3), 1242–1255.

Lively, S. E., Pisoni, D. B., Yamada, R. A., Yoh'ichi, T., & Yamada, T. (1994).
Training Japanese listeners to identify english /r/ and /l/. III. Long-term retention of new phonetic categories. *Journal of the Acoustical Society of America*, *96*(4), 2076–2087. https://doi.org/10.1121/1.410149

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese Listeners To Identify English /R/ And /1/: A First Report. *Journal of the Acoustical Society of America*, *89*(2), 874–886. https://doi.org/10.1121/1.1894649

Logan, J. S., & Pruitt, J. S. (1995). Methodological issues in training listeners to perceive non-native phonemes. *Speech Perception and Linguistic Experience. Issues in Cross-Language Research*, 351–377.

Mahboob, A., & Elyas, T. (2014). English in the Kingdom of Saudi Arabia. *World Englishes*, *33*(1), 128–142. https://doi.org/10.1111/weng.12073

Matsuura, H. (2007). Intelligibility and individual learner differences in the EIL context. *System*, *35*(3), 293–304. https://doi.org/10.1016/J.SYSTEM.2007.03.003

McArthur, T. (1998). *The English languages*. Cambridge University Press.

McCandliss, B. D., Fiez, J. A., Protopapas, A., Conway, M., & McClelland, J. L. (2002). Success and failure in teaching the [r]-[l] contrast to Japanese adults: Tests of a Hebbian model of plasticity and stabilization in spoken language perception. *Cognitive, Affective and Behavioral Neuroscience*, *2*(2), 89–108. https://doi.org/10.3758/CABN.2.2.89

McMahon, A. M. (2002). *An introduction to English phonology*. Edinburgh University Press.

Mitchell, T. F. (1993). *Pronouncing Arabic 2*. Oxford: Clarendon Press.

Mitleb, F. (1984). Vowel length contrast in Arabic and English: a spectrographic test. *Journal of Phonetics*, *12*(3), 229–235.

Morrison, G. S., & Assmann, P. (2013). Vowel inherent spectral change. In *Vowel Inherent Spectral Change*. Springer Science and Business Media.

Morrow, P. R. (2004). English in Japan: the world Englishes perspective. *JALT Journal*.

Mousa, A. (1994). *The interphonolgy of Saudi learners of English*. PhD thesis, University of Essex.

Munro, M. J. (1993). Productions of English vowels by native speakers of Arabic: Acoustic measurements and accentedness ratings. *Language and Speech*, *36*(1), 39–66. https://doi.org/10.1177/002383099303600103

Munro, M. J., & Derwing, T. M. (1995). Foreign Accent, Comprehensibility, and Intelligibility in the Speech of Second Language Learners. *Language Learning*, *45*(1), 73–97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x

Munro, M. J., & Derwing, T. M. (2006). The functional load principle in ESL pronunciation instruction: An exploratory study. *System*, *34*(4), 520–531. https://doi.org/10.1016/J.SYSTEM.2006.09.004

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*(3), 313–334.

NEOM. (2020). *[NEOM Scholarship Program] برنامج نيوم للابتعاث*. https://sr.neom.com/ar/education/programmes/neom-scholarship-preporatory-program

Newman, D. L. (2002). The phonetic status of Arabic within the worlds languages: the uniqueness of the lughat al-daad. *Antwerp Papers in Linguistics*, *100*, 65–75.

Newman, D., & Verhoeven, J. (2002). Frequency analysis of Arabic vowels in connected speech. *Antwerp Papers in Linguistics*, *100*, 77–86.

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

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204–238.

Nozawa, T. (2015). Effects of training methods and attention on the identification and discrimination of American English coda nasals by native Japanese listeners. *The Journal of the Acoustical Society of America*, *138*(3), 1947–1947. https://doi.org/10.1121/1.4934161

Olson, D. J. (2014). Phonetics and technology in the classroom: A practical approach to using speech analysis software in second-language pronunciation instruction. *Hispania*, *97*(1), 47–68. https://doi.org/10.1353/hpn.2014.0030

Peterson, G. E., & Barney, H. L. (1952). Control Methods Used in a Study of the Vowels. *The Journal of the Acoustical Society of America*, *24*(2), 175–184. https://doi.org/10.1121/1.1906875

Pickering, L. (2006). Current research on intelligibility in english as a lingua franca. *Annual Review of Applied Linguistics*, *26*, 219–233.

Pisoni, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception & Psychophysics*, *13*(2), 253–260.

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.

Rajadurai, J. (2007). Intelligibility studies: a consideration of empirical and ideological issues. *World Englishes*, *26*(1), 87–98.

Rampton, M. B. H. (1990). Displacing the 'native speaker': expertise, affiliation, and inheritance. *ELT Journal*, *44*(2), 97–101. https://doi.org/10.1093/ELTJ/44.2.97

Rato, A. A. D. S. (2014). *Cross-language Perception and Production of English Vowels by Portuguese Learners: The Effects of Perceptual Training*.

Reetz, H., & Jongman, A. (2020). *Phonetics: Transcription, Production, Acoustics and Perception*. John Wiley & Sons.

Roach, P. (2010). *English Phonetics and Phonology: A practical course* (4th ed.). Cambridge University Press.

Sadakata, M., & McQueen, J. M. (2013). High stimulus variability in nonnative speech learning supports formation of abstract categories: Evidence from Japanese geminates. *The Journal of the Acoustical Society of America*, *134*(2), 1324–1335.

Saudi Vision 2030. (2023). *[Saudi Vision 2030] 2030 رؤية السعودية*. https://www.vision2030.gov.sa/en/

Schneider, C. (2011). Why Field Linguists Should Pay More Attention to Research in Applied Linguistics. *Australian Journal of Linguistics*, *31*(2), 187–209.

Shang, S., Nesson, E., & Fan, M. (2018). Interaction Terms in Poisson and Log Linear Regression Models. *Bulletin of Economic Research*, *70*(1), E89–E96.

Sheldon, A., & Strange, W. (1982). The acquisition of /r/ and /l/ by Japanese learners of English: Evidence that speech production can precede speech perception. *Applied Psycholinguistics*, *3*(3), 243–261.

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

Smith, L. E., & Nelson, C. L. (1985). International intelligibility of English: directions and resources. *World Englishes*, *4*(3), 333–342. https://doi.org/10.1111/J.1467-971X.1985.TB00423.X

Strange, W., & Dittmann, S. (1984). Effects of discrimination training on the perception of /r-l/ by Japanese adults learning English. *Perception & Psychophysics*, *36*(2), 131–145. https://doi.org/10.3758/BF03202673

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

Tersta, F. W., & Novianti, A. (2017). *Listening to Students Voice:Students' Problems in Listening Comprehension*.

Thomson, R. I. (2018). High Variability [Pronunciation] Training (HVPT). *Journal of Second Language Pronunciation*, *4*(2), 208–231. https://doi.org/10.1075/jslp.17038.tho

Timmis, I. (2007). *The attitudes of language learners towards target varieties of the language*. Language acquisition and development.

Trudgil, P. (2003). *A Glossary of Sociolinguistics*. Oxford University Press.

Uchihara, T., Karas, M., & Thomson, R. (2021). *High Variability Phonetic Training (HVPT): A meta-analysis.*

Ur, P. (1996). *A Course in English Language Teaching*. Cambridge University Press.

Wadifa.com. (2023). *الهيئة العامة للترفيه تعلن ٥ وظائف ادارية وهندسية لحملة البكالوريوس في الرياض [The General Entertainment Authority announces 5 administrative and engineering jobs for bachelor’s degree holders in Riyadh]*. https://www.wzufa.com/job/general-entertainment-authority-september-2023/

Wang, X., & Munro, M. J. (2004). Computer-based training for learning English vowel contrasts. *System*, *32*(4), 539–552. https://doi.org/10.1016/J.SYSTEM.2004.09.011

Watson, J. (2002). *The phonology and morphology of Arabic*. Oxford University Press.

Wayland, R. P., & Li, B. (2008). Effects of two training procedures in cross-language perception of tones. *Journal of Phonetics*, *36*(2), 250–267.

Wells, J. C. (1982a). *Accents of English* (Vol. 1). Cambridge University Press.

Wells, J. C. (1982b). *Accents of English 2: The British Isles*. Cambridge 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., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, *7*(1), 49–63.

Willis, J. (1996). A Framework for Task-Based Learning. In *Creative Education*. Longman.

Willoughby, L., & Manns, H. (2019). *Australian English reimagined: structure, features and developments*. Routledge.

Wong, J. W. S. (2012). Training the Perception and Production of English /e/ and /æ/ of Cantonese ESL Learners: A Comparison of Low vs. High Variability Phonetic Training. *Proceedings of the 14th Australasian International Conference on Speech Science and Techniology*, 37–40.

Wong, J. W. S. (2014). The effects of high and low variability phonetic training on the perception and production of english vowels /e/-/æ/ by cantonese ESL learners with high and low l2 proficiency levels. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *August 2013*, 524–528.

Yamada, R. A., & Tohkura, Y. (1992). The effects of experimental variables on the perception of American English /r/ and /l/ by Japanese listeners. *Perception & Psychophysics*, *52*(4), 376–392.

Yavas, M. (2006). *Applied English Phonology*. John Wiley & Sons.

Yavas, M. (2020). *Applied English Phonology*. John Wiley & Sons.

Zhang, X., Cheng, B., & Zhang, Y. (2021). The Role of Talker Variability in Nonnative Phonetic Learning: A Systematic Review and Meta-Analysis. *Journal of Speech, Language, and Hearing Research*, *64*(12), 4802–4825. https://doi.org/10.1044/2021_JSLHR-21-00181

Zimmermann, A. G. (1993). *Inference about the fixed and random effects in a mixed-effects linear model: an approximate Bayesian approach*. Iowa State University.

Appendices

Appendix A. Participant information sheet

Newcastle University
School of Education, Communication & Language Sciences

You are invited to take part in a research study. Before you decide whether or not you wish to take part, it is important that you understand why the research is being done and what it will involve. Please read this information carefully and discuss it with others if you wish. Take time to decide whether or not you wish to take part. The study is conducted by Sarah Alghabban as part of her PhD studies at Newcastle University and is supervised by Dr Ghada Khattab and Dr Jalal Al-Tamimi.

The research involves a three-month phonetics training program; thus, if you want to participate, you should plan to complete the training. The training will take place on the university's campus. If you agree to take part in this study, you will be asked to sign an informed consent form. Regardless, you can withdraw at any moment without incurring any obligations. You will be asked to complete 16 sessions of a phonetics training study in order to improve your listening to and understanding of English. The training contains two sessions per week and each session will last for 50 minutes. You will also be asked to complete pre-, mid- and post- tests with the total time for completing each test not exceeding 35 minutes. Throughout the training and testing sessions, you will listen to stimuli containing all English vowels and carry out a set of tasks which enable you to better identify and produce these vowels by multiple speakers. You will also be recorded reading a word list (during training sessions) and a word list plus a short story (during the testing phase). Prior to the training, you will be asked to complete a short questionnaire about your language background.

All responses you give, or other data collected will be anonymised and saved on a secure password-protected drive. In any research report that may be published, no information will be included that will make it possible to identify you individually. There will be no way to connect your name to your responses at any time during or after the study.

Newcastle University is the sponsor for this study, based in the United Kingdom. Newcastle University will be using information from you in order to undertake this study and will act as the data controller for this study. This means that Newcastle University is responsible for looking after your information and using it properly.

Your rights to access, change or move your information are limited, as Newcastle University need to manage your information in specific ways in order for the research to be reliable and accurate. If you withdraw from the study, Newcastle University will keep the information about you that has already been obtained. To safeguard your rights, the minimum personally identifiable information will be used.

If you have any questions or concerns regarding this project or wish to withdraw, please email the researcher at B6053991@newcastle. ac.uk

Appendix B. Language background questionnaire

Date:

Below are questions about your background, language history, language use, education, and beliefs about accents. Please answer these questions as completely as possible.

Section 1: Background

Age:

Sex:

* Were you born in Saudi Arabia (SA)? **Yes** **No**

If no:

How old were you when you came to SA?

* Have you lived in SA since birth? **Yes**
No

If no:

How long have you been living in SA?

* Have you ever lived in an English-speaking country (e.g., UK, US)? (if yes, for how long)? **Yes** **No**

Section 2: Language History

* What is your native language?
* What is your father's native language?
* What is your mother's native language?

Please list any other languages that you know below. For each, rate how well you can use the language on the following scale:

Not Good 1 2 3 4 5 Very Good					
Other language(s)	Pronunciation	Listening	Speaking	Reading	Grammar
1					
2					
3					

For the languages you listed, including your native language, please indicate below the place and age at which you learned them, and if applicable, whether you learned them by formal lessons (e.g., at school or a course), or by informal learning (e.g., at home, at work, from friends).

Language	Country	Age	Lessons (yes/no)	Duration of lessons	Informal (yes/no)	Duration of informal learning
1 (native language)						
2						
3						

For the languages you listed, including your native language, which do you use with the following people, for how many hours per day.

Individuals	Language 1	hr/day	Language 2	hr/day	Language 3	hr/day
Parents						
Siblings						
Other family members						
Partners						
Friends						
Colleages						
Others (please specify)						

For the languages you listed, which do you use for the following activities and for how many hours per day?

Activity	Language 1	hr/day	Language 2	hr/day	Language 3	hr/day
Watching TV						
Listening to the radio/podcasts/ music						
Meeting people online						
Reading						
Writing						
Email, internet						

In general, how well do you *like* to learn new languages?

Dislike 1 2 3 4 5 **Like**

In general, how *easy* do you find learning new languages?

Difficult 1 2 3 4 5 **Easy**

If you have any other comments about your language history that you think may be important for your ability to use these languages, please feel free to write them here:

____

____

____

____

____

____

____

____

____

Section 3: **Education-related information**

* What academic degree are you pursuing? and what level are you at now?

____

*Have you been taught by L1 English speakers? **Yes** **No**

*Have you been taught by L2 English instructors from SA? **Yes** **No**

*Have you been taught by L2 English instructors other than Saudis? **Yes** **No**

If yes:

What the L1 the instructors speak other than English or Saudi Arabic?

a) ___________________

b) ___________________

c) ___________________

d) ___________________

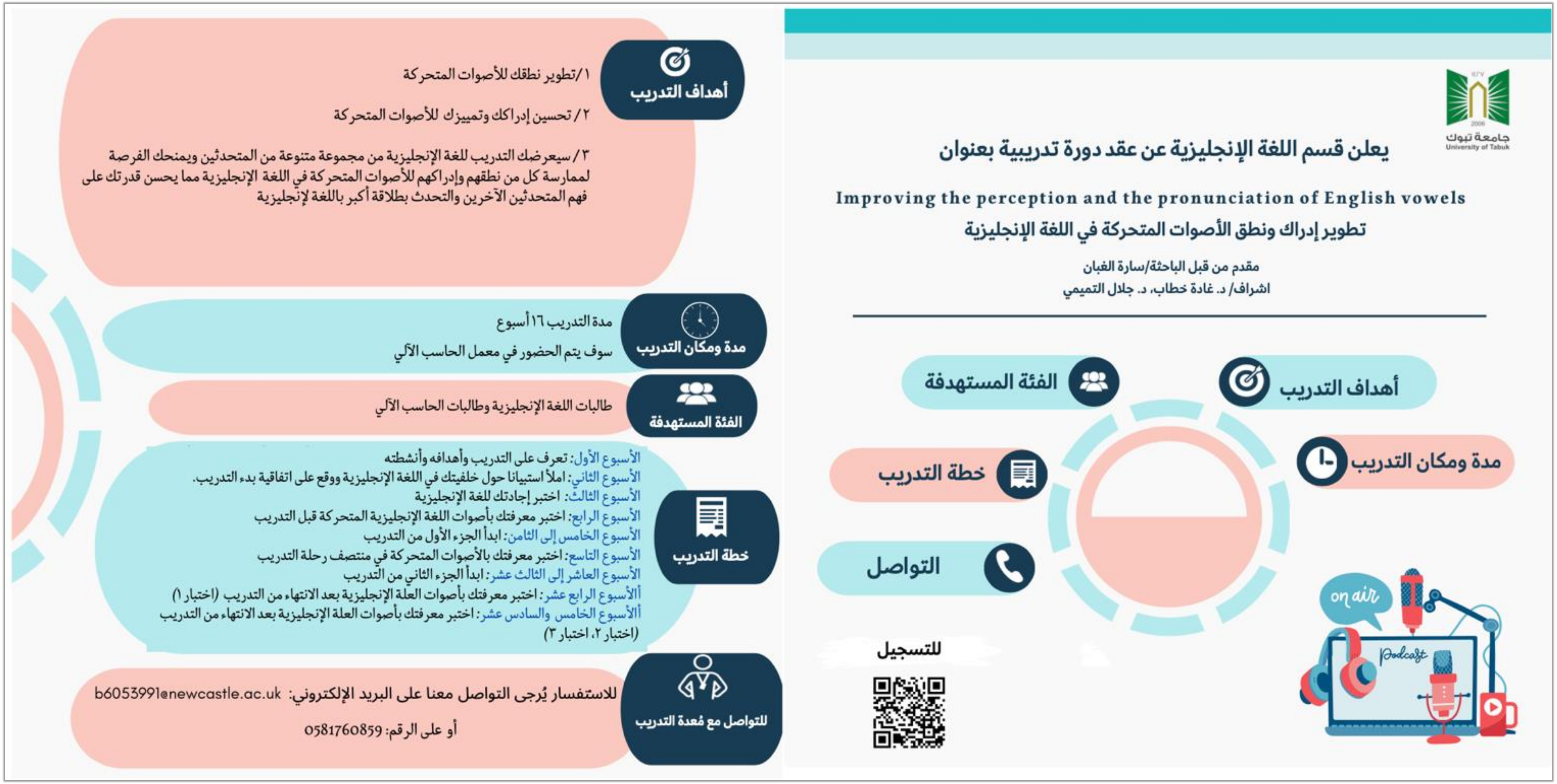
Section 4: **Beliefs and points of view about accents**

The following statements reflect beliefs about L1 and L2 English accents. Please indicate your opinion by clicking on one of the numbers between 1 and 5.

	Strongly Disagree				**Strongly Agree**
1. Maintaining my Saudi Arabic accent while speaking English English reflects who I am and how I want to be introduced.	**1**	**2**	**3**	**4**	**5**
2. I spend a great deal of time and effort replicating/imitating the sound of an L1 speaker (e.g., a British, American)	**1**	**2**	**3**	**4**	**5**
3. When L1 or L2 English speakers converse with me, I believe they understand me well.	**1**	**2**	**3**	**4**	**5**
4. I am satisfied with my present proficiency in English pronunciation and identification of vowels	**1**	**2**	**3**	**4**	**5**
5. I'm looking to improve my English pronunciation and Identification of vowels	**1**	**2**	**3**	**4**	**5**

Appendix C. The advertisement poster for the training course

C.I Arabic Version



جامعة تبوك
University of Tabuk

يعلن قسم اللغة الإنجليزية عن عقد دورة تدريبية بعنوان

Improving the perception and the pronunciation of English vowels

تطوير إدراك ونطق الأصوات المتحركة في اللغة الإنجليزية

مقدم من قبل الباحثة/سارة الغبان
اشراف/ د. غادة خطاب، د. جلال التميمي

أهداف التدريب

الفئة المستهدفة

مدة ومكان التدريب

خطة التدريب

التواصل

للتسجيل

on air
Podcast

أهداف التدريب

١/تطوير نطقك للأصوات المتحركة

٢/ تحسين إدراكك وتمييزك للأصوات المتحركة

٣/ سيعرضك التدريب للغة الإنجليزية من مجموعة متنوعة من المتحدثين ويمنحك الفرصة لممارسة كل من نطقهم وإدراكهم للأصوات المتحركة في اللغة الإنجليزية مما يحسن قدرتك على فهم المتحدثين الآخرين والتحدث بطلاقة أكبر باللغة لإنجليزية

مدة ومكان التدريب

مدة التدريب ١٦ أسبوع

سوف يتم الحضور في معمل الحاسب الآلي

الفئة المستهدفة

طالبات اللغة الإنجليزية وطالبات الحاسب الآلي

خطة التدريب

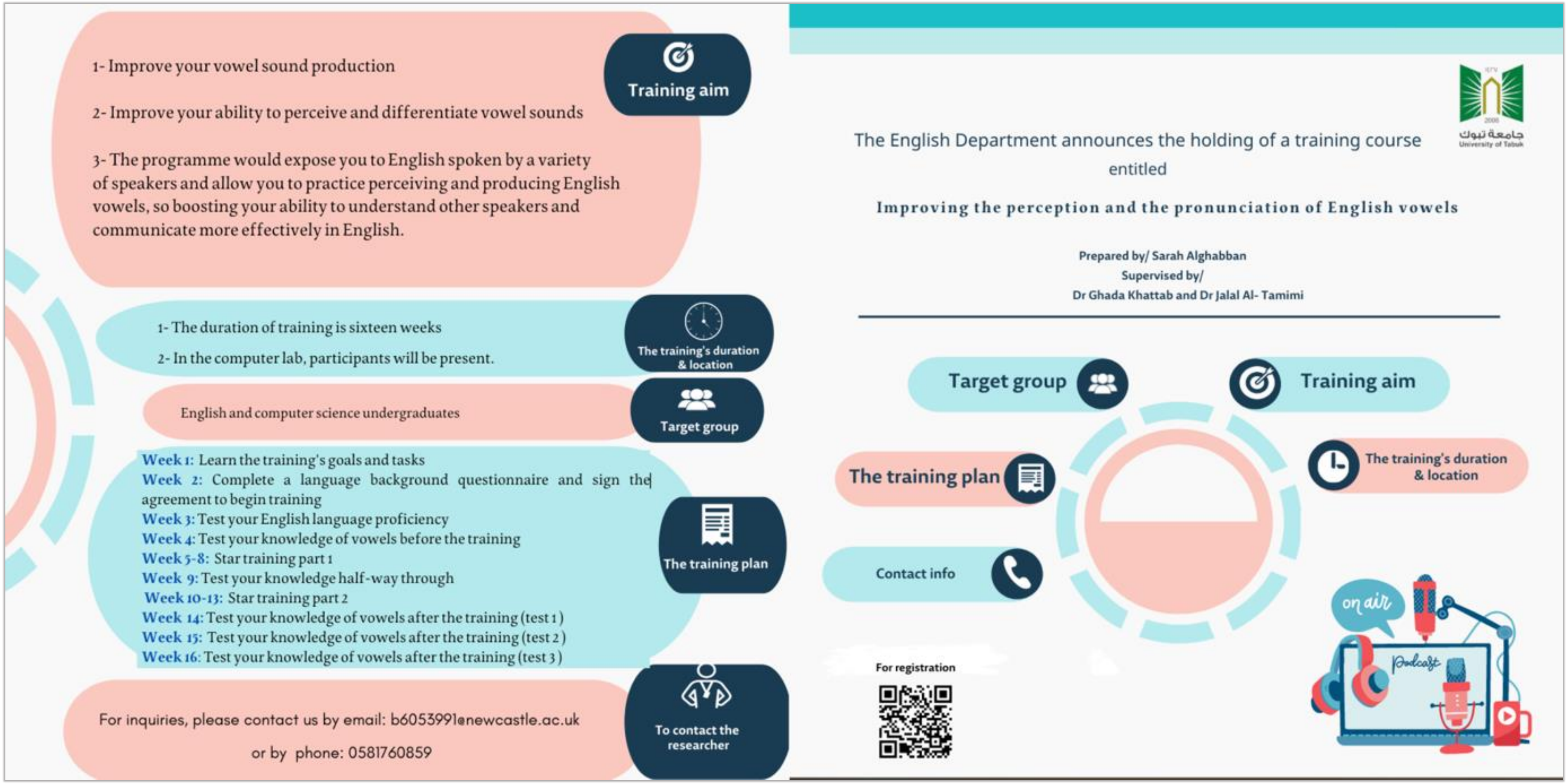
الأسبوع الأول: تعرف على التدريب وأهدافه وأنشطته
الأسبوع الثاني: املأ استبيانا حول خلفيتك في اللغة الإنجليزية ووقع على اتفاقية بدء التدريب.
الأسبوع الثالث: اختبر إجادتك للغة الإنجليزية
الأسبوع الرابع: اختبر معرفتك بأصوات اللغة الإنجليزية المتحركة قبل التدريب
الأسبوع الخامس إلى الثامن: ابدأ الجزء الأول من التدريب
الأسبوع التاسع: اختبر معرفتك بالأصوات المتحركة في منتصف رحلة التدريب
الأسبوع العاشر إلى الثالث عشر: ابدأ الجزء الثاني من التدريب
أالأسبوع الرابع عشر: اختبر معرفتك بأصوات العلة الإنجليزية بعد الانتهاء من التدريب (اختبار ١)
أالأسبوع الخامس والسادس عشر: اختبر معرفتك بأصوات العلة الإنجليزية بعد الانتهاء من التدريب (اختبار ٢، اختبار ٣)

للتواصل مع مُعدة التدريب

للاستفسار يُرجى التواصل معنا على البريد الإلكتروني: b6053991@newcastle.ac.uk

أو على الرقم: 0581760859

C.II English Version



Training aim

1- Improve your vowel sound production

2- Improve your ability to perceive and differentiate vowel sounds

3- The programme would expose you to English spoken by a variety of speakers and allow you to practice perceiving and producing English vowels, so boosting your ability to understand other speakers and communicate more effectively in English.

The training's duration & location

1- The duration of training is sixteen weeks

2- In the computer lab, participants will be present.

Target group

English and computer science undergraduates

The training plan

Week 1: Learn the training's goals and tasks
Week 2: Complete a language background questionnaire and sign the agreement to begin training
Week 3: Test your English language proficiency
Week 4: Test your knowledge of vowels before the training
Week 5-8: Star training part 1
Week 9: Test your knowledge half-way through
Week 10-13: Star training part 2
Week 14: Test your knowledge of vowels after the training (test 1)
Week 15: Test your knowledge of vowels after the training (test 2)
Week 16: Test your knowledge of vowels after the training (test 3)

To contact the researcher

For inquiries, please contact us by email: b6053991@newcastle.ac.uk

or by phone: 0581760859

جامعة تبوك
University of Tabuk

The English Department announces the holding of a training course entitled

Improving the perception and the pronunciation of English vowels

Prepared by/ Sarah Alghabban
Supervised by/
Dr Ghada Khattab and Dr Jalal Al- Tamimi

Target group

Training aim

The training plan

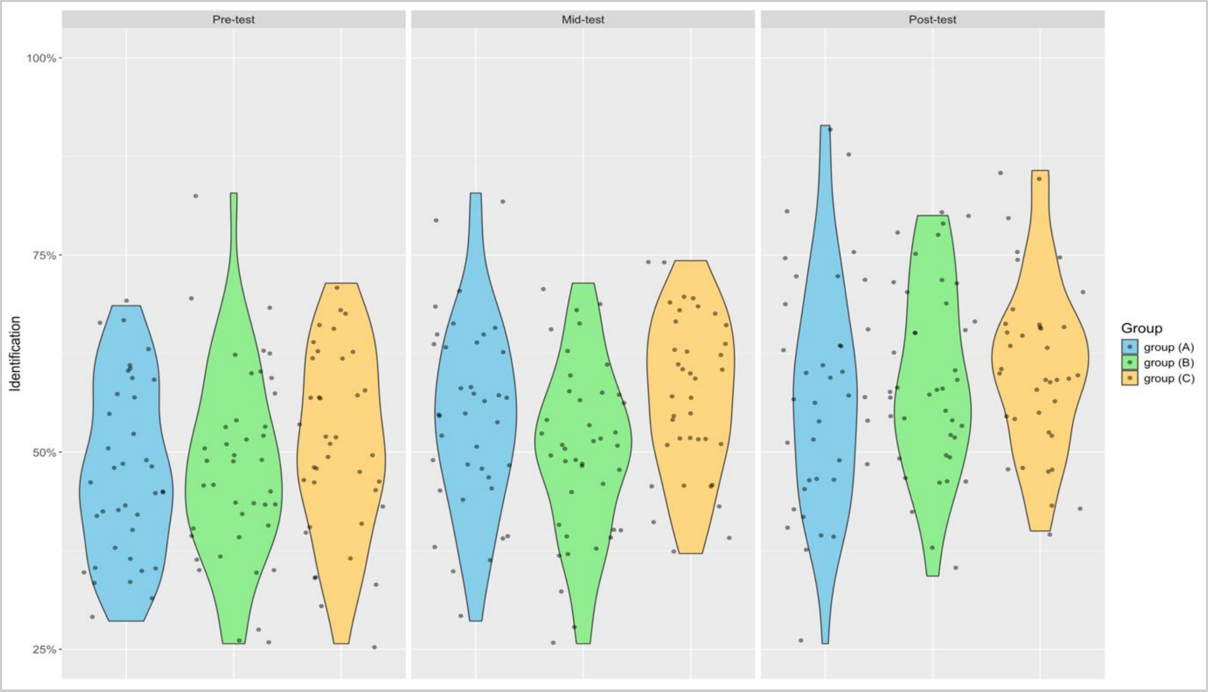
The training's duration & location

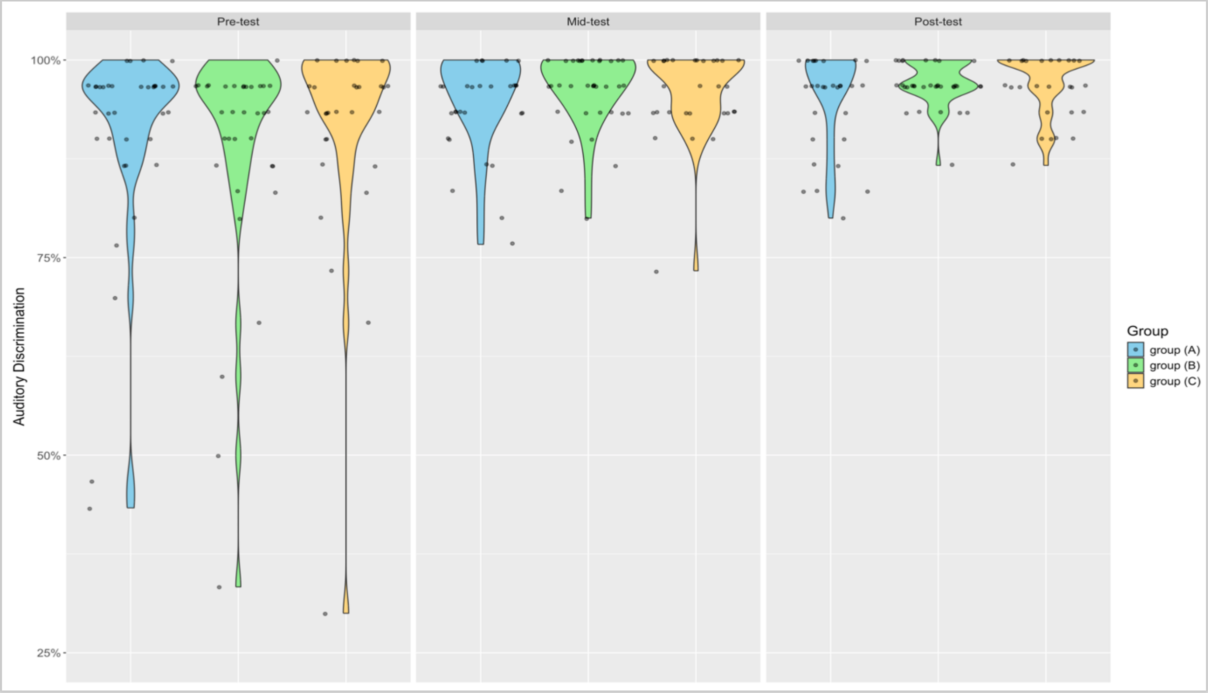
Contact info

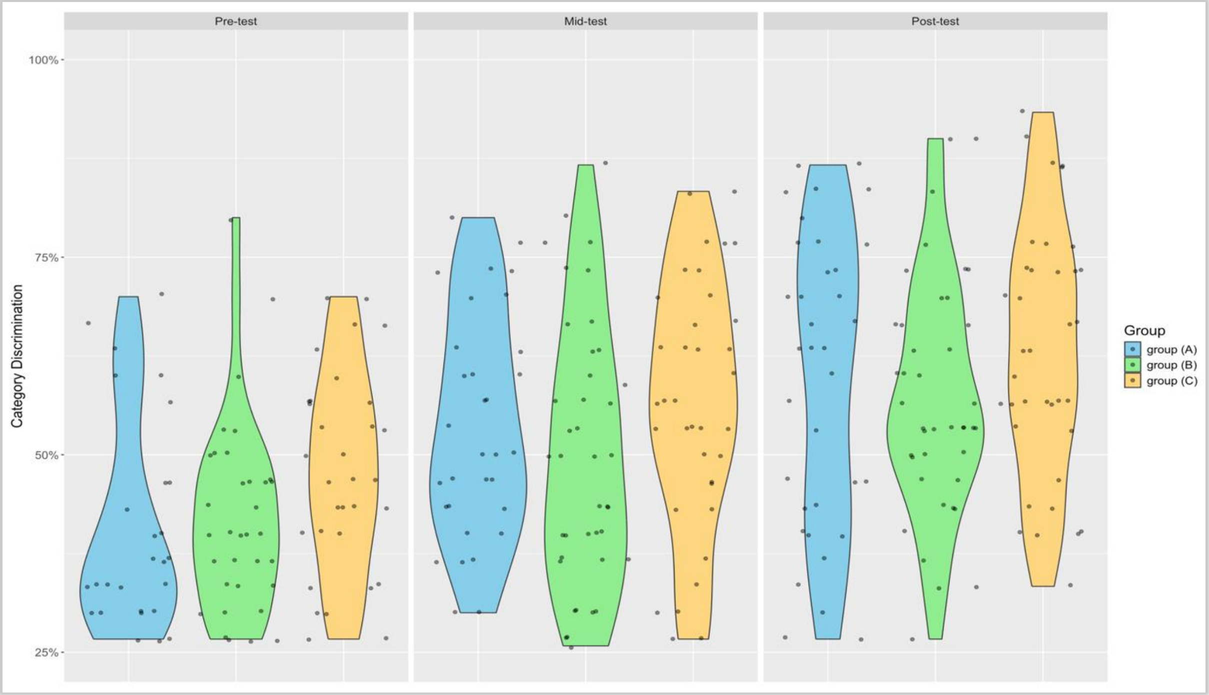
For registration

Appendix D. Raw data visualisations

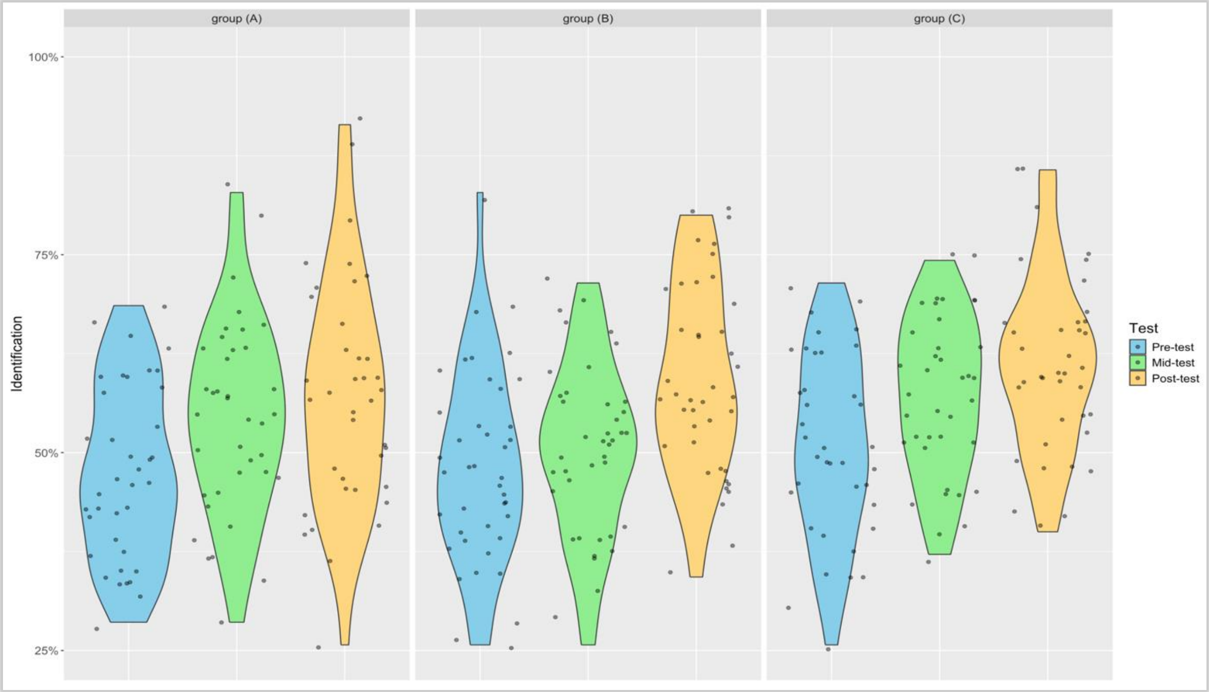
D.I Groups (A, B, C) across pre-, mid-, and post-tests for identification, auditory discrimination and category discrimination tasks

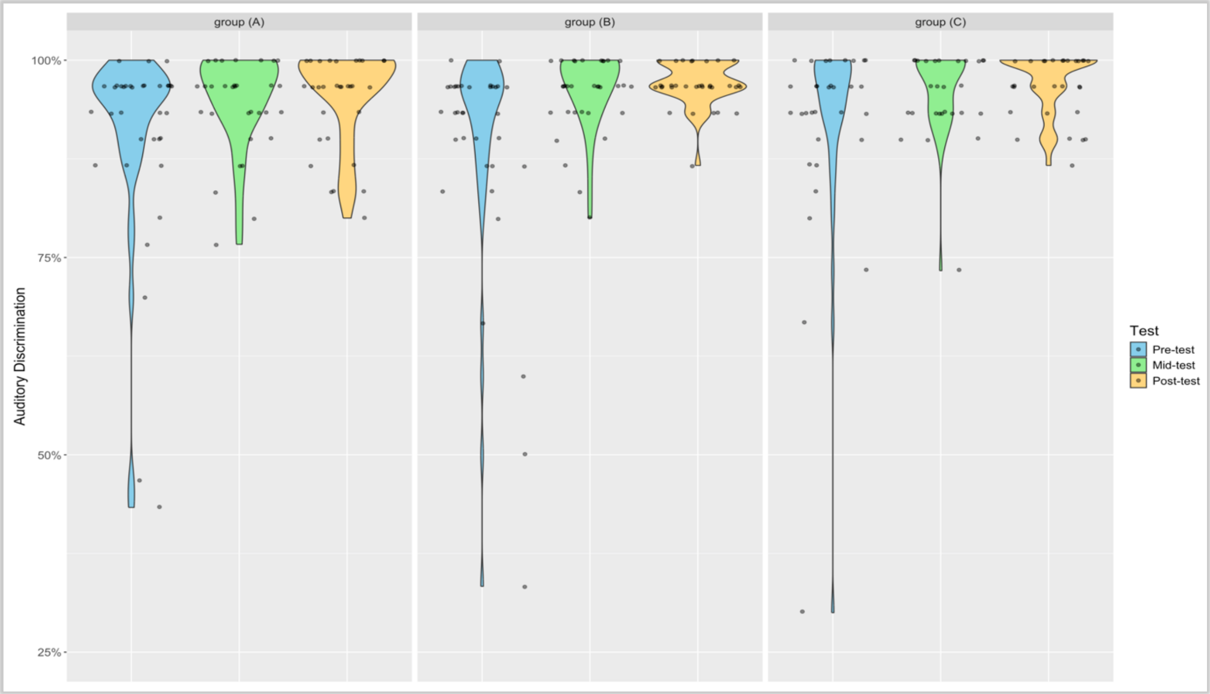


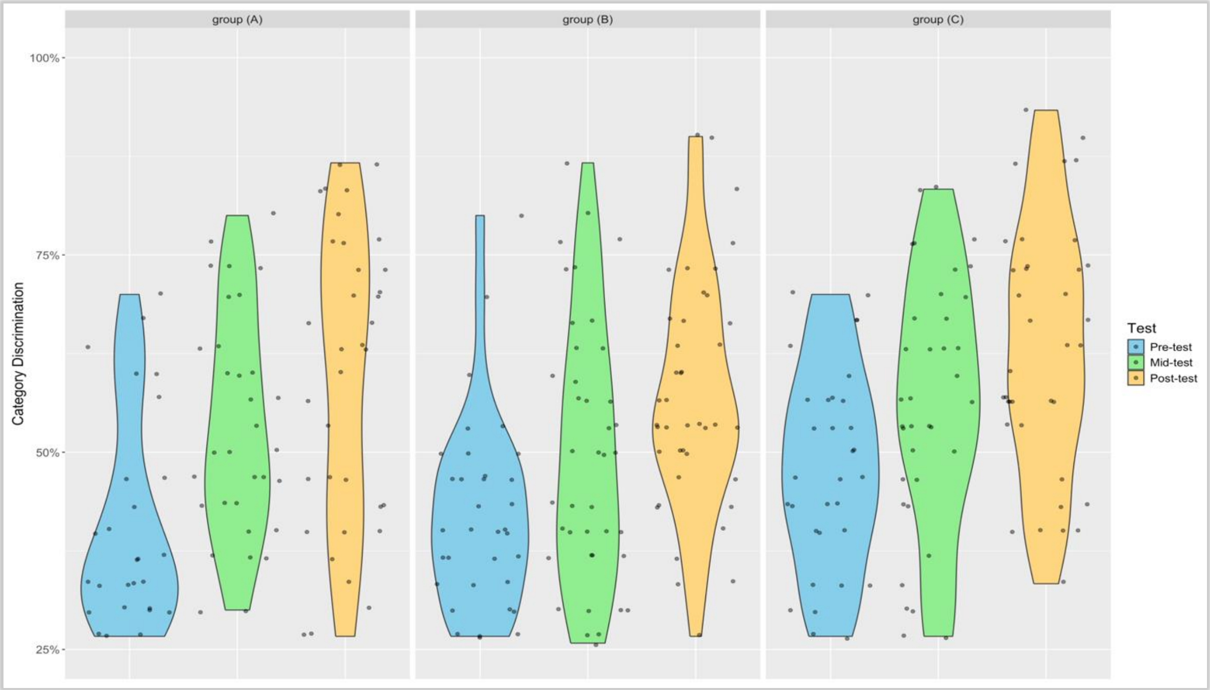




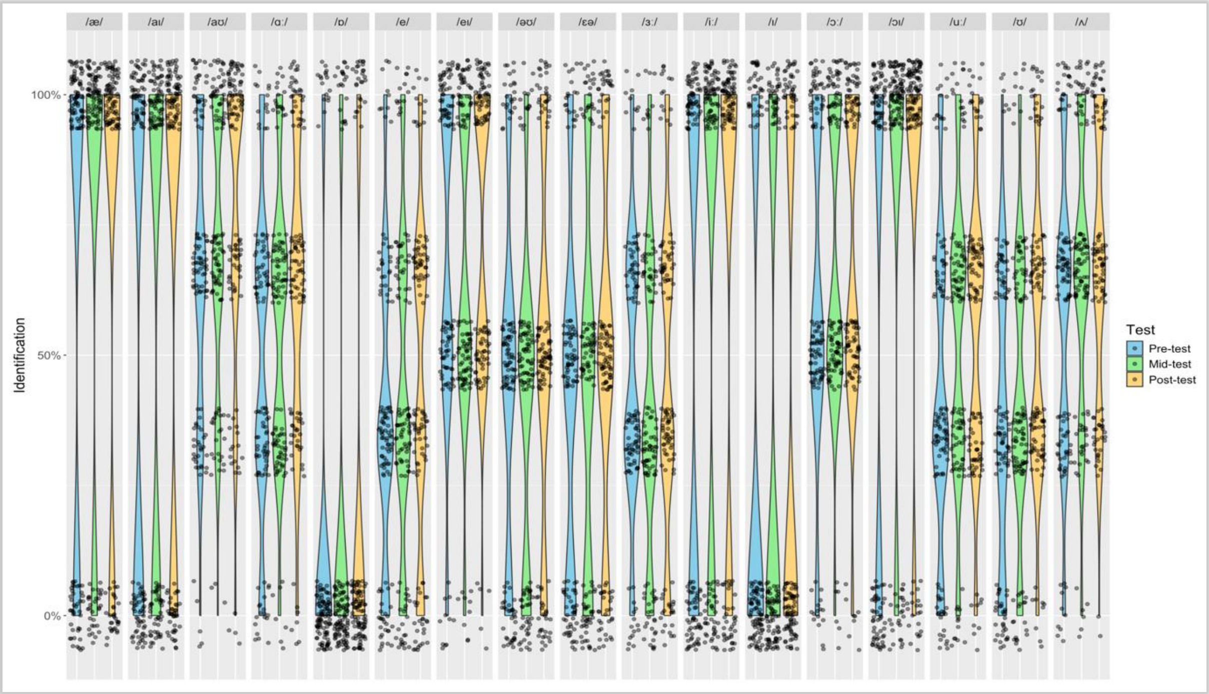
D.II Tests (pre-, mid-, and post-) across groups A, B, and C for identification, auditory discrimination and category discrimination tasks

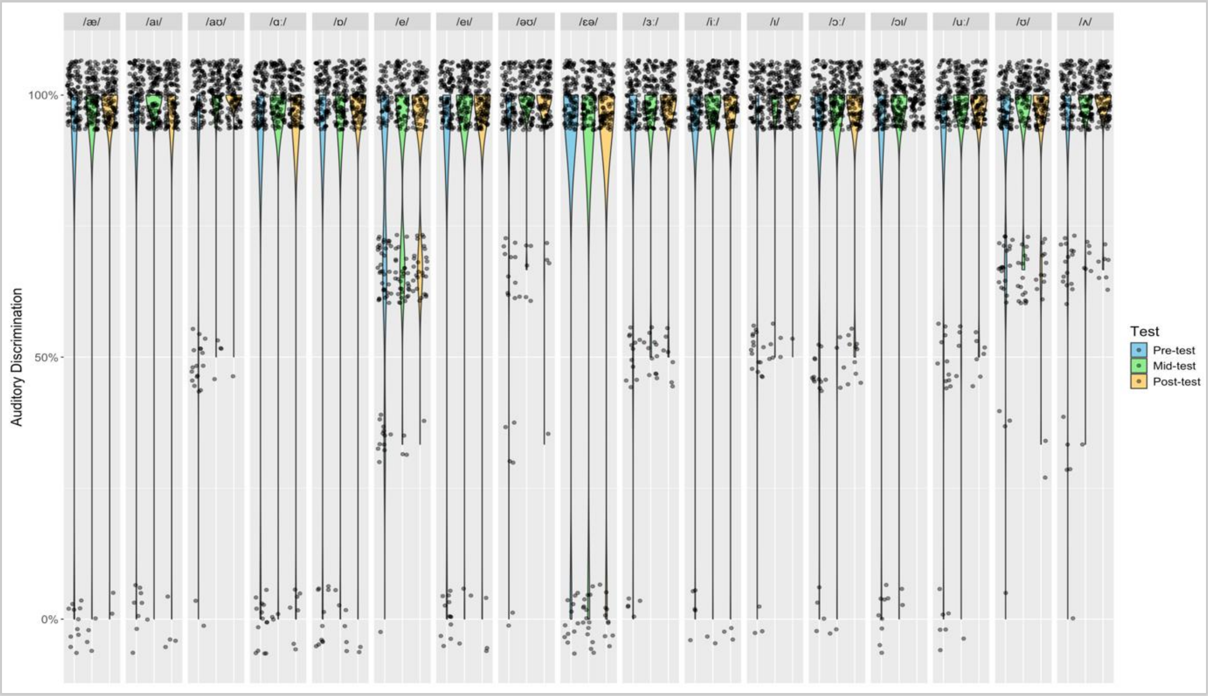


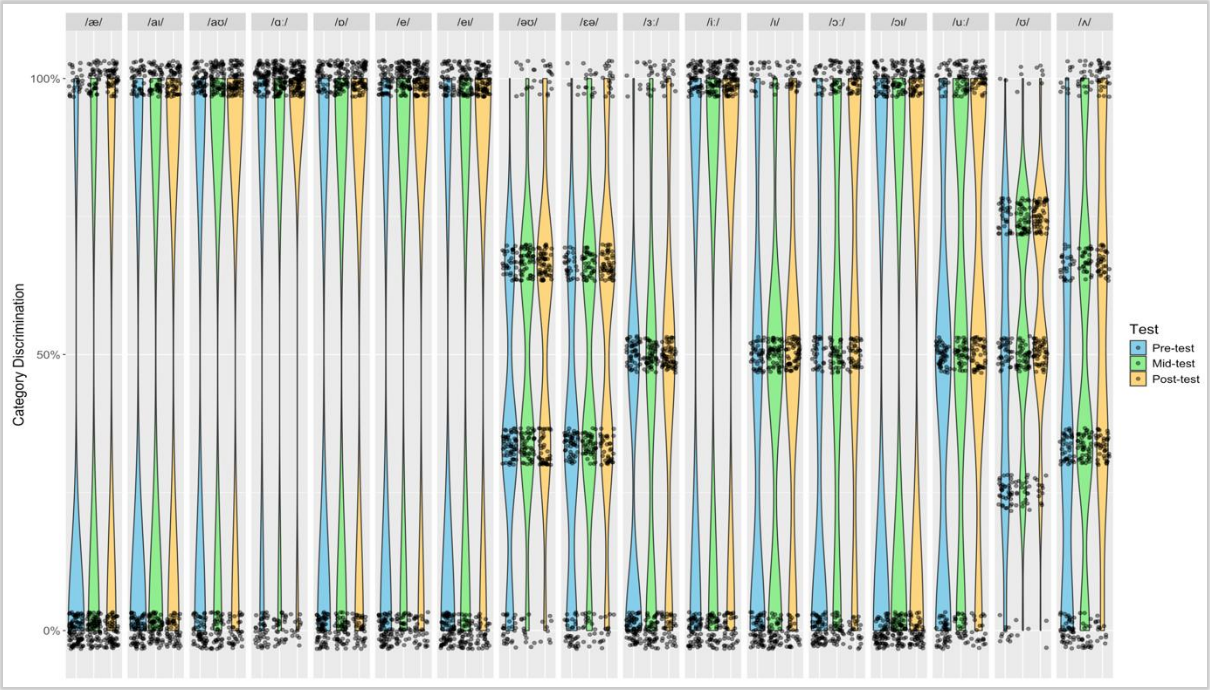

group (A)
group (B)
group (C)
Auditory Discrimination
100%
75%
50%
25%
Test
Pre-test
Mid-test
Post-test


group (A)
group (B)
group (C)
Category Discrimination
100%
75%
50%
25%
Test
Pre-test
Mid-test
Post-test

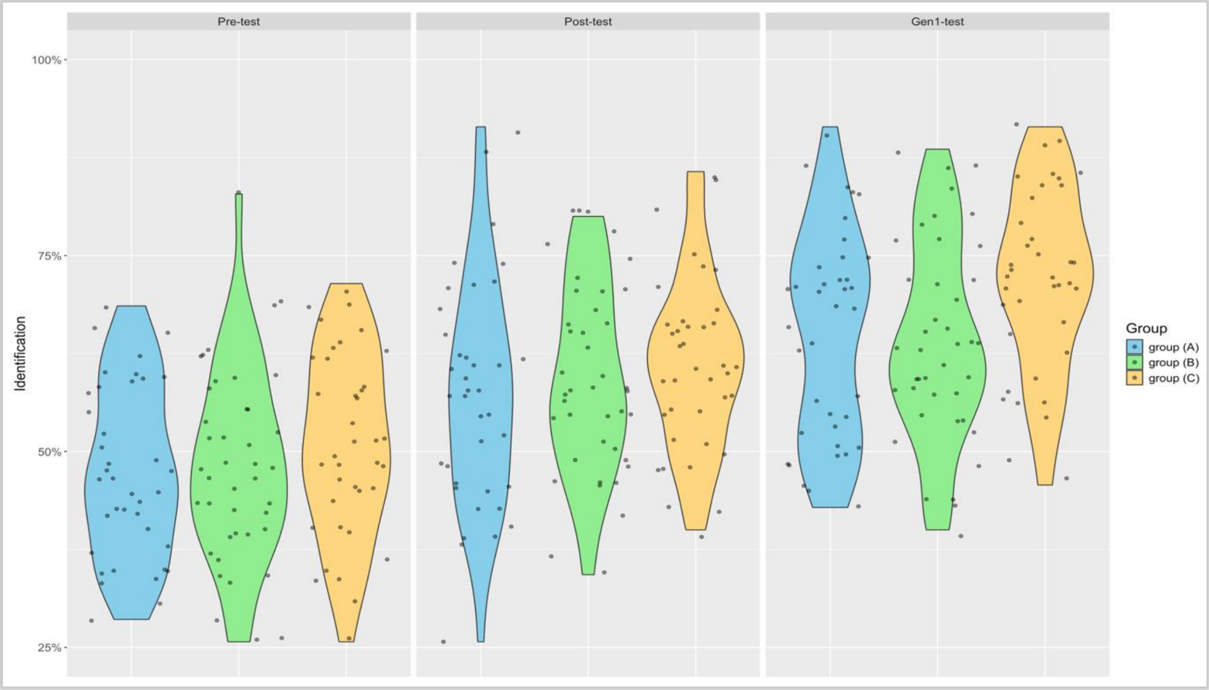
D.III Vowels across pre-, mid-, and post-tests for identification, auditory discrimination and category discrimination tasks

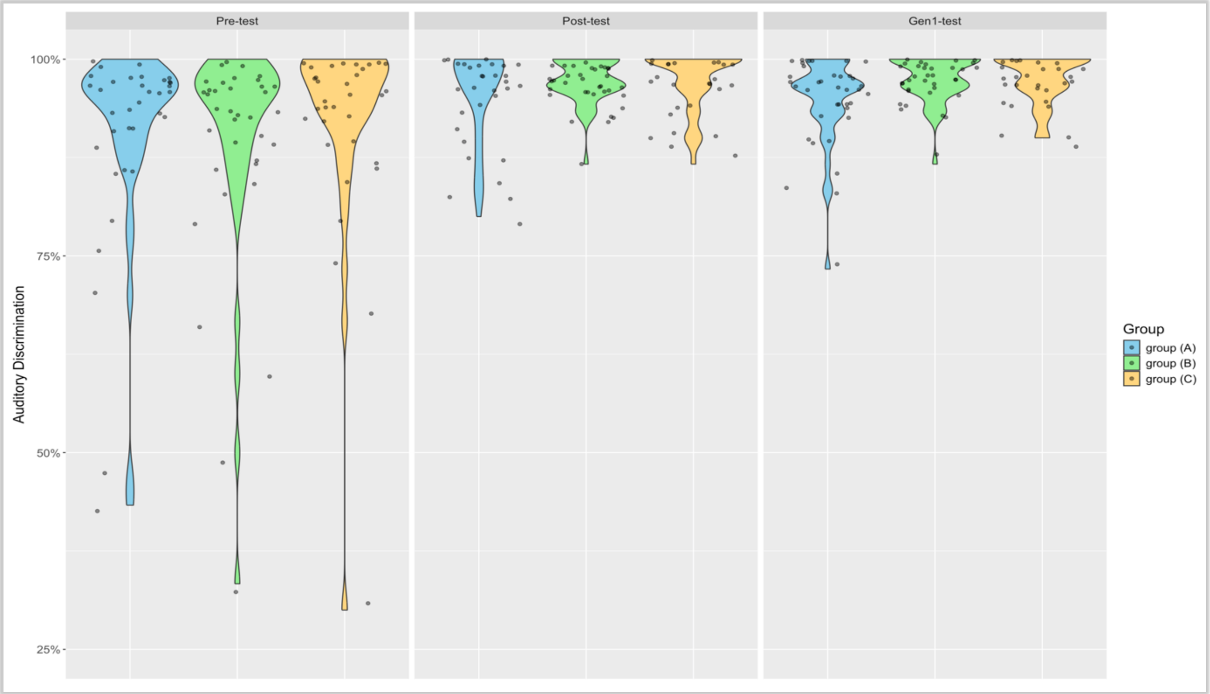


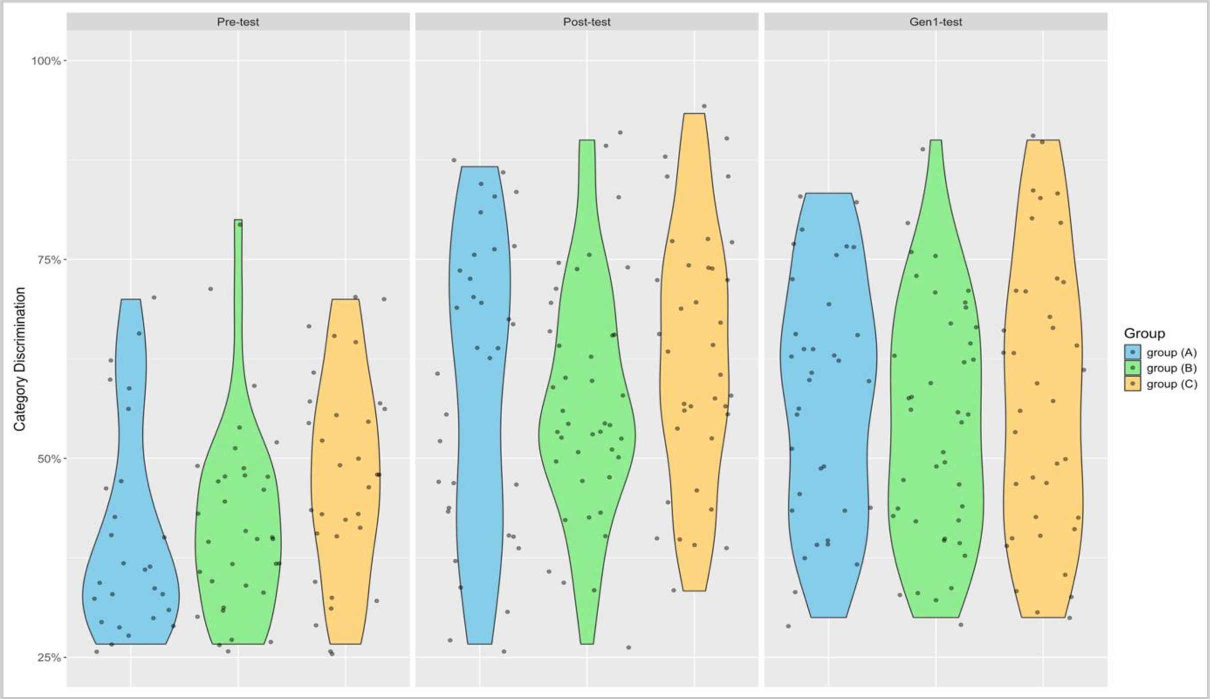




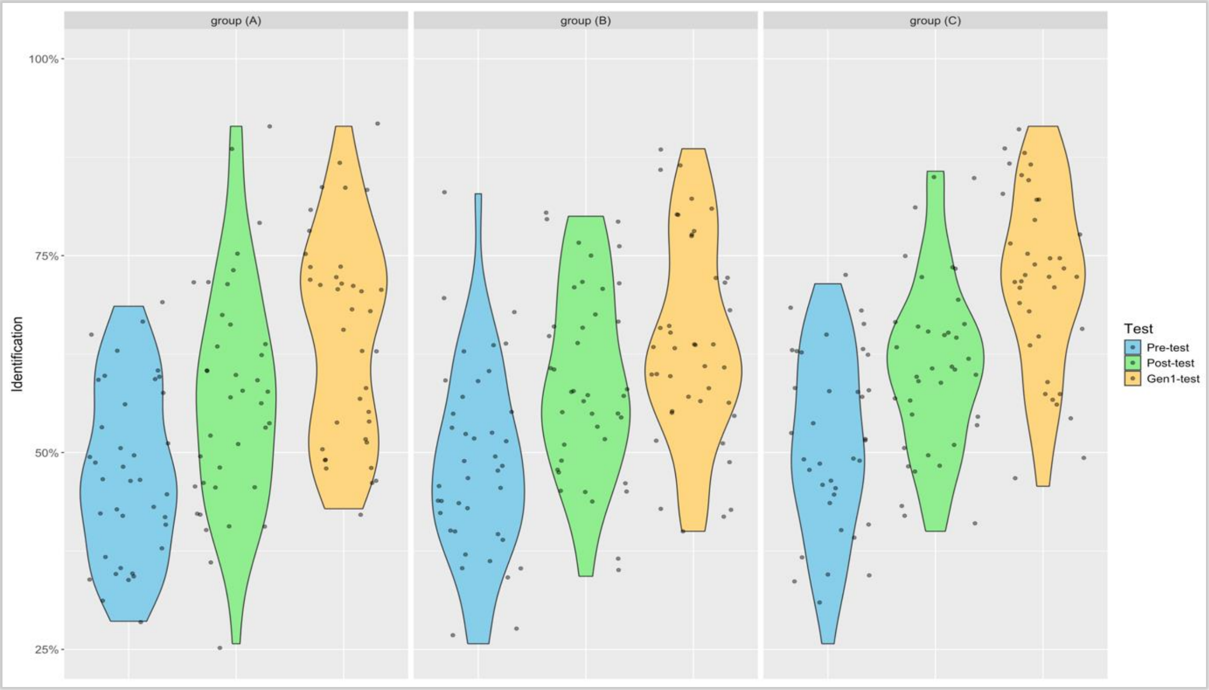
D.IV Groups (A, B, C) across gen1, post- and pre- tests for identification, auditory discrimination and category discrimination tasks

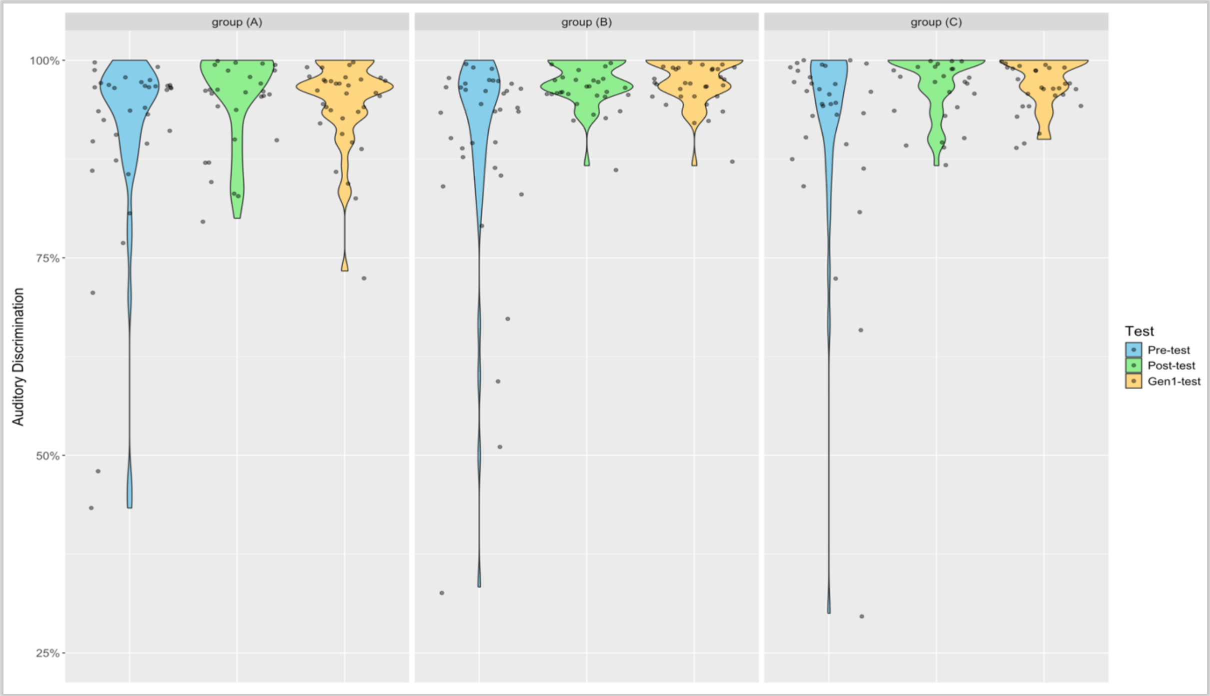


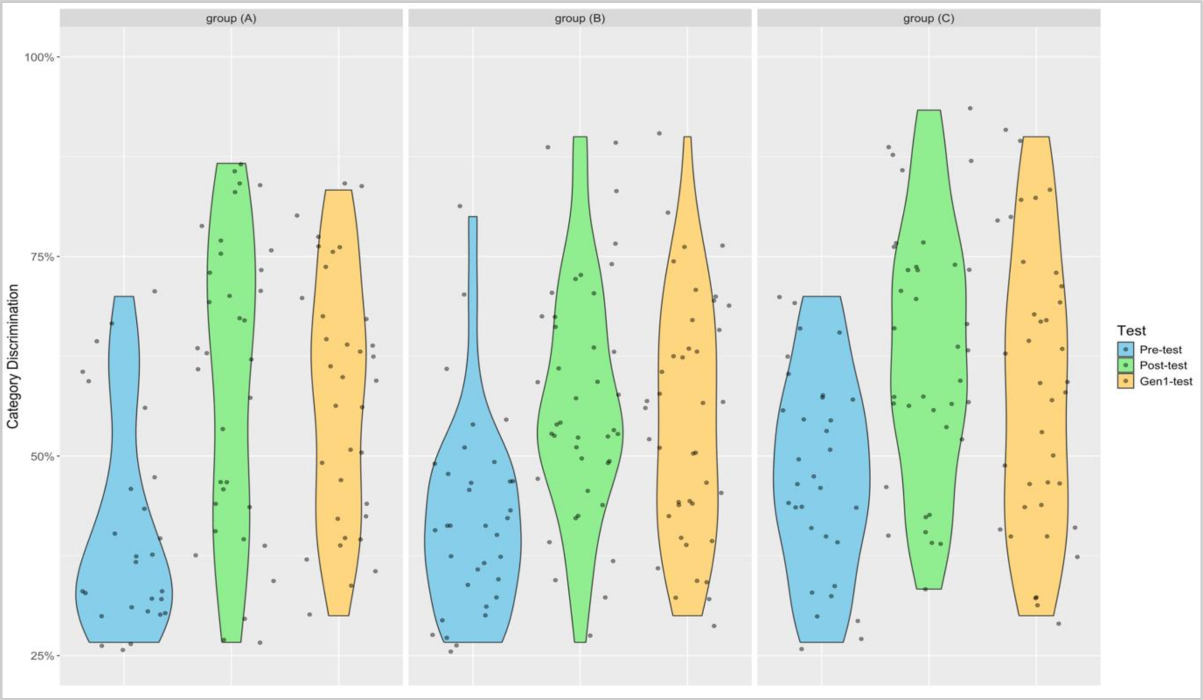

Pre-test
Post-test
Gen1-test
Auditory Discrimination
100%
75%
50%
25%
Group
group (A)
group (B)
group (C)


Pre-test
Post-test
Gen1-test
Category Discrimination
100%
75%
50%
25%
Group
group (A)
group (B)
group (C)

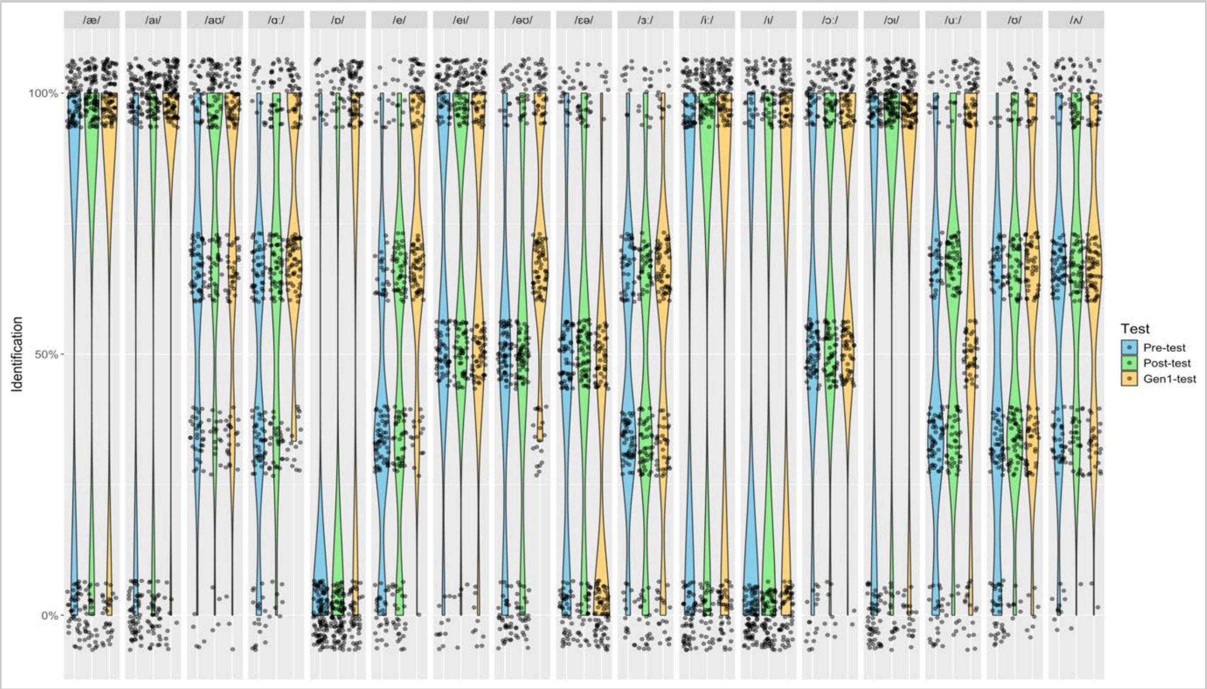
D.V Tests (gen1, post-, and pre-) across groups A, B, and C for identification, auditory discrimination and category discrimination tasks

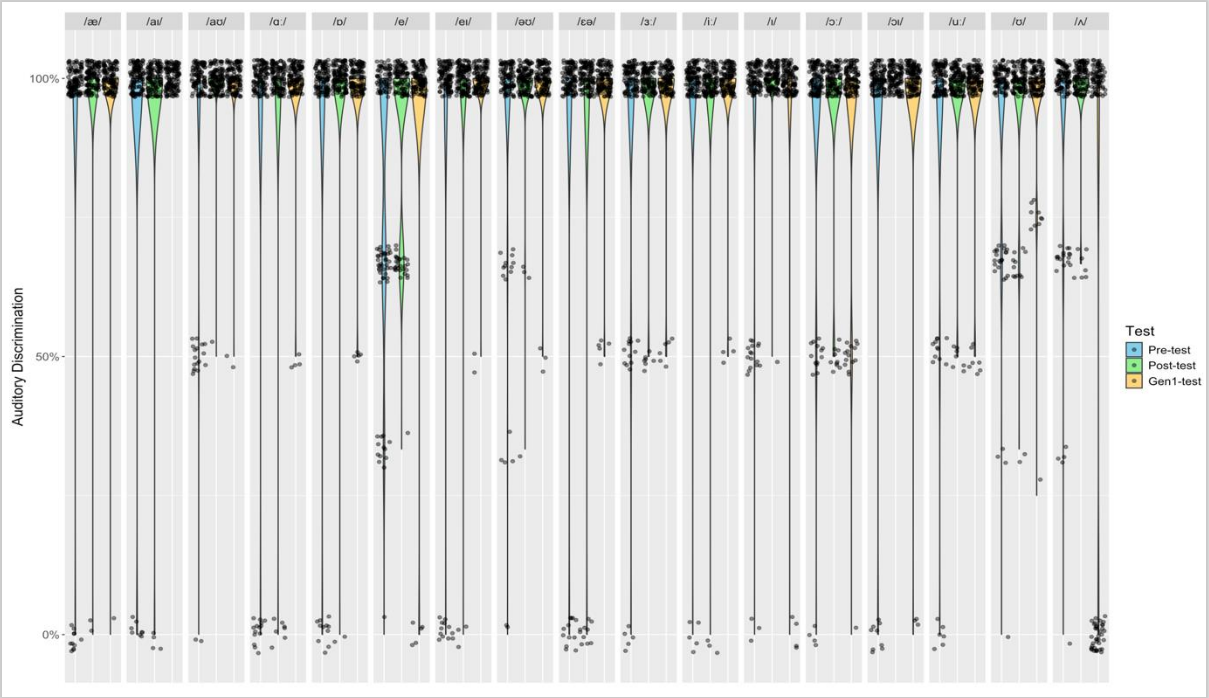


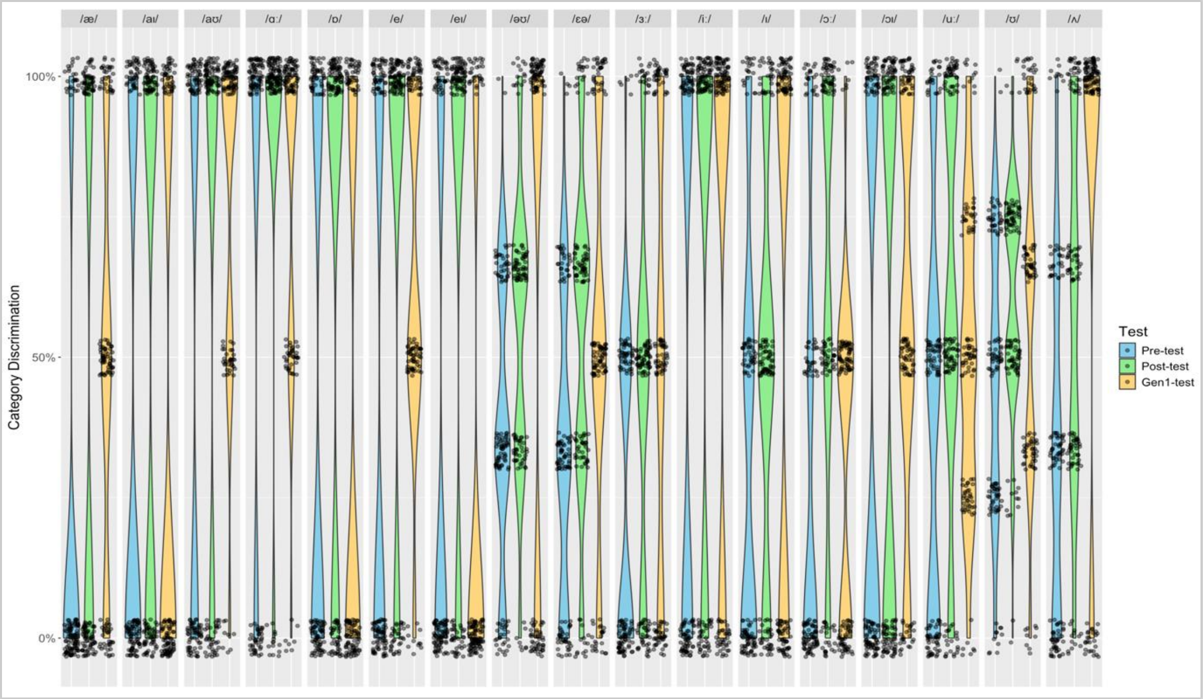




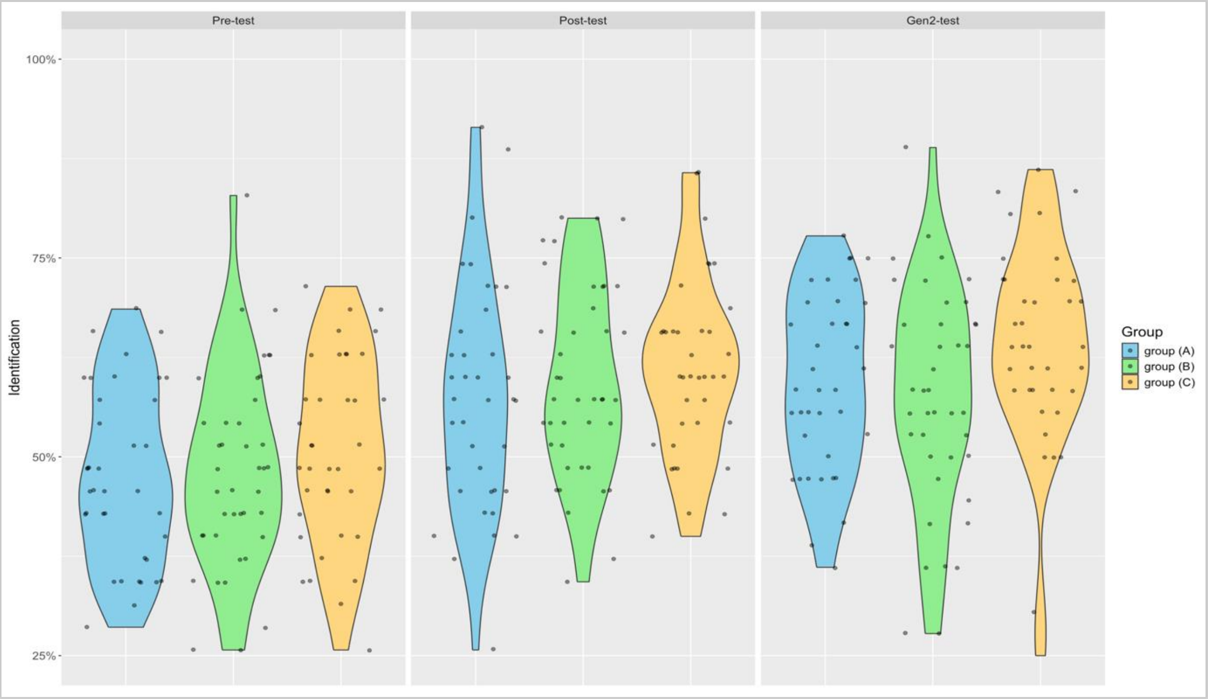
V.I Vowels across gen1, post-, and pre-tests for identification, auditory discrimination and category discrimination tasks

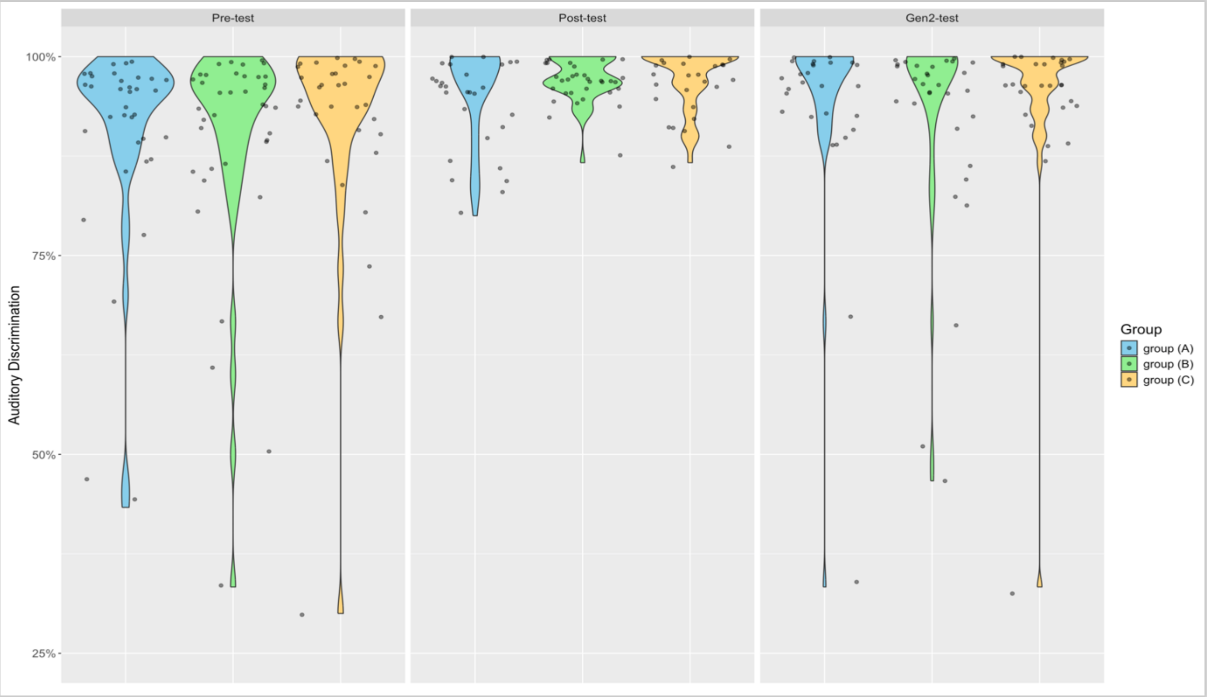



/æ/
/aɪ/
/aʊ/
/ɑ:/
/ɒ/
/e/
/eɪ/
/əʊ/
/ɛə/
/ɜ:/
/i:/
/ɪ/
/ɔ:/
/ɔɪ/
/u:/
/ʊ/
/ʌ/
100%
50%
0%
Auditory Discrimination
Test
Pre-test
Post-test
Gen1-test

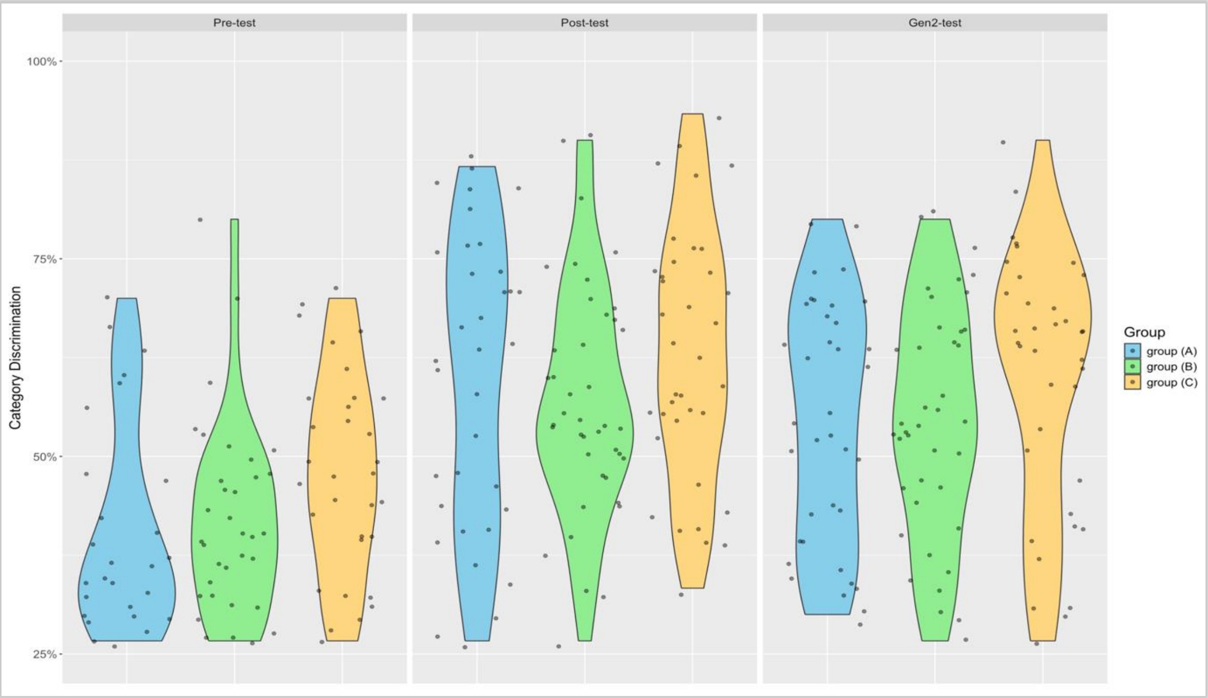

/æ/
/aɪ/
/aʊ/
/ɑ:/
/ɒ/
/e/
/eɪ/
/əʊ/
/ɛə/
/ɜ:/
/i:/
/ɪ/
/ɔ:/
/ɔɪ/
/u:/
/ʊ/
/ʌ/
100%
50%
0%
Category Discrimination
Test
Pre-test
Post-test
Gen1-test

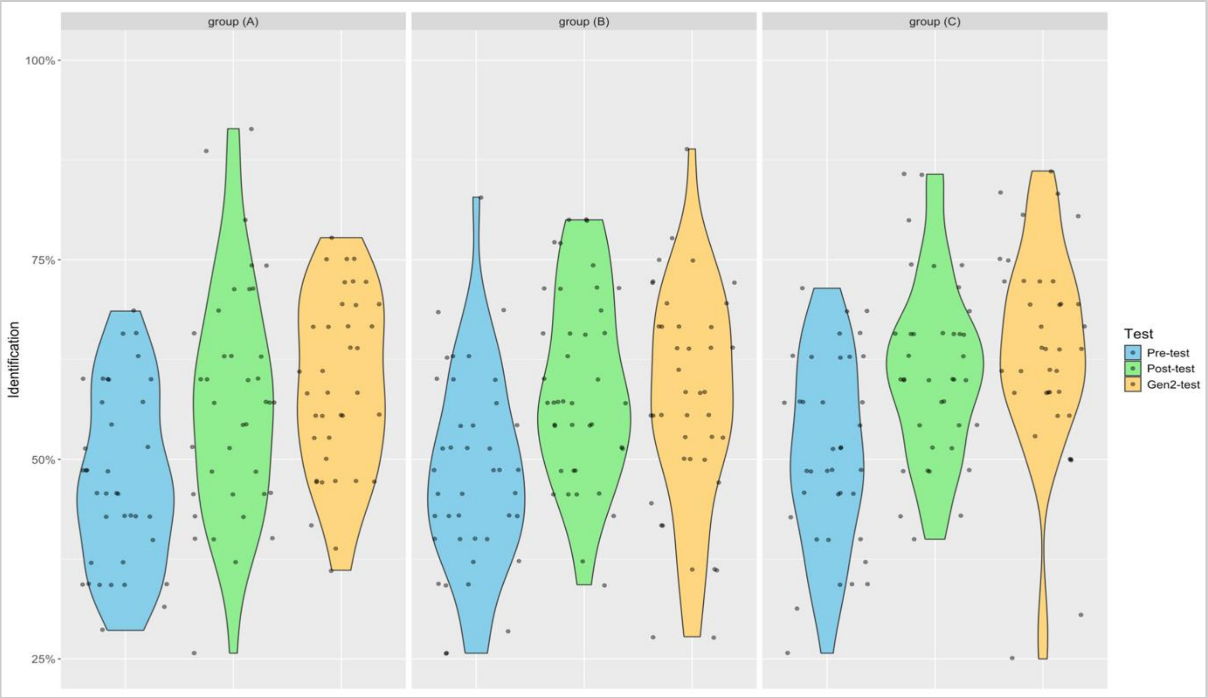
D.VII Groups (A, B, C) across gen2, post- and pre- tests for identification, auditory discrimination and category discrimination tasks

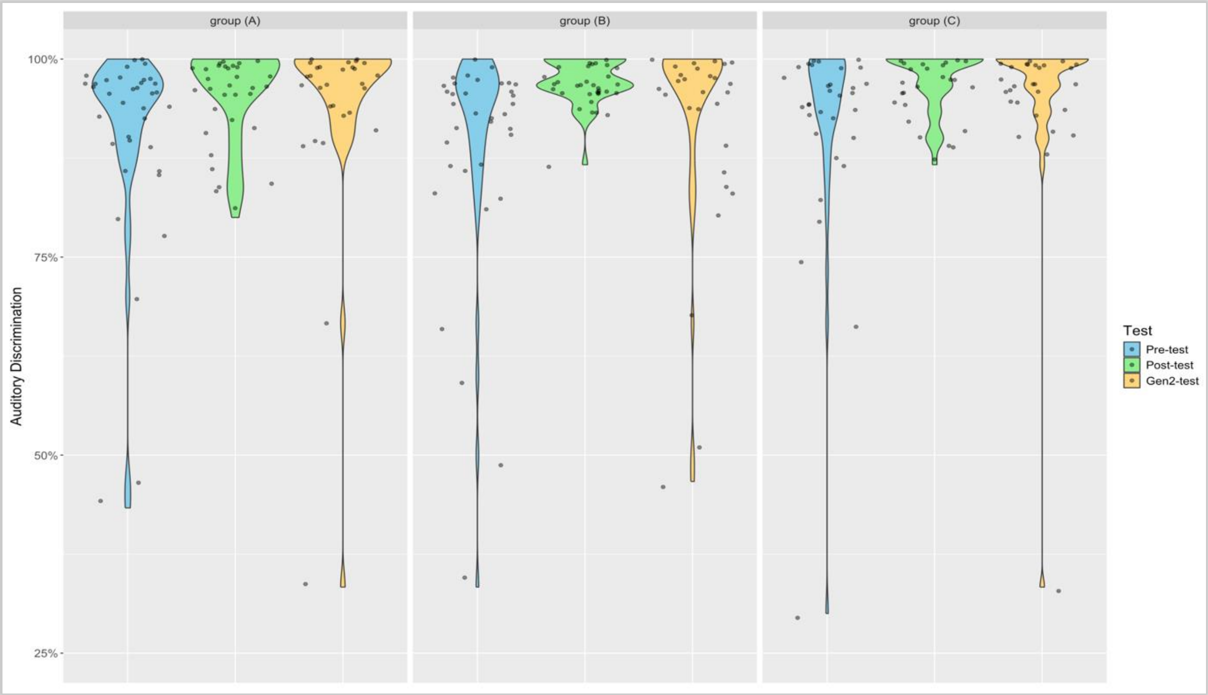




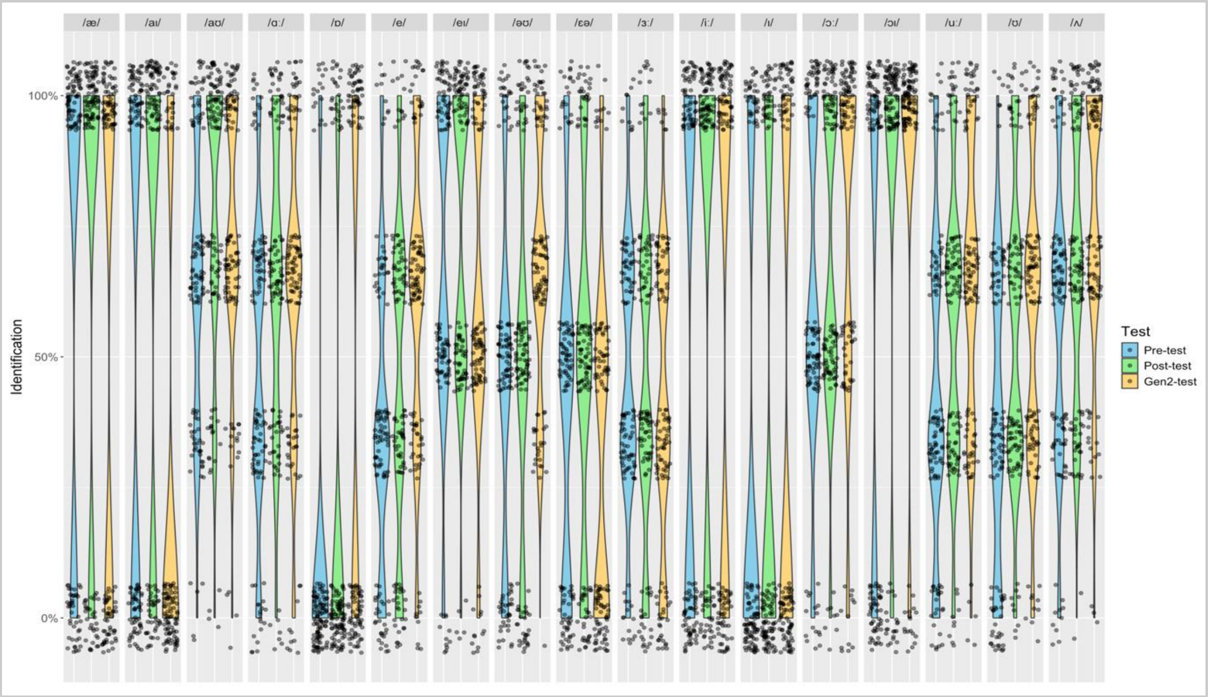
D.VIII Tests (gen2, post-, and pre-) across groups A, B, and C for identification, auditory discrimination and category discrimination tasks

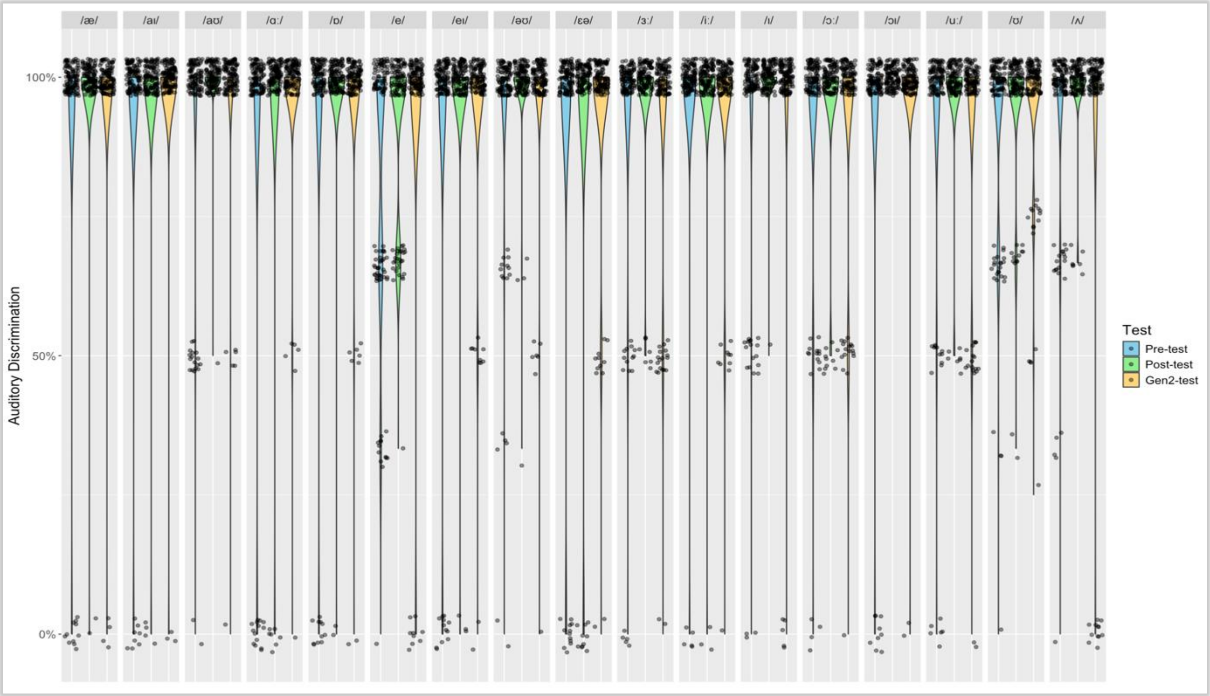


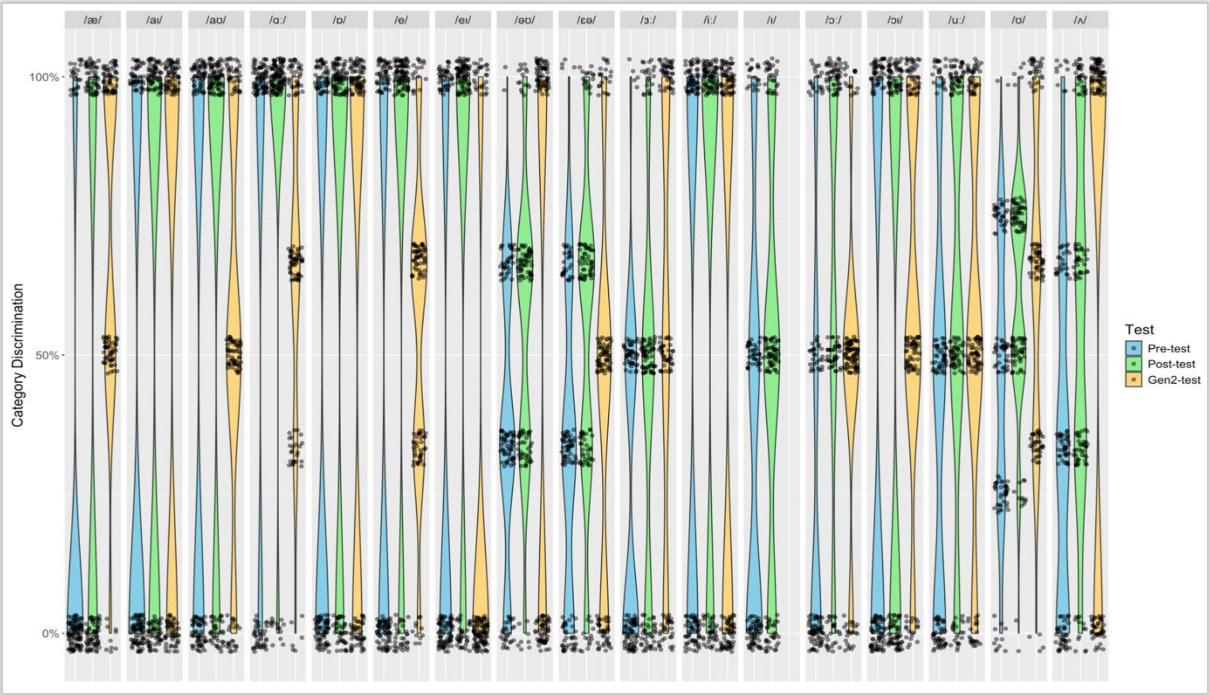




I.X Vowels across gen2, post-, and pre-tests for identification, auditory discrimination and category discrimination tasks




/æ/
/aɪ/
/aʊ/
/ɑː/
/ɒ/
/e/
/eɪ/
/əʊ/
/ɛə/
/ɜː/
/iː/
/ɪ/
/ɔː/
/ɔɪ/
/uː/
/ʊ/
/ʌ/
100%
50%
0%
Auditory Discrimination
Test
Pre-test
Post-test
Gen2-test


/æ/
/aɪ/
/aʊ/
/ɑː/
/ɒ/
/e/
/eɪ/
/əʊ/
/ɛə/
/ɜː/
/iː/
/ɪ/
/ɔː/
/ɔɪ/
/uː/
/ʊ/
/ʌ/
100%
50%
0%
Category Discrimination
Test
Pre-test
Post-test
Gen2-test

Appendix E. Testing stimuli

E.I pre-/mid-/post-test

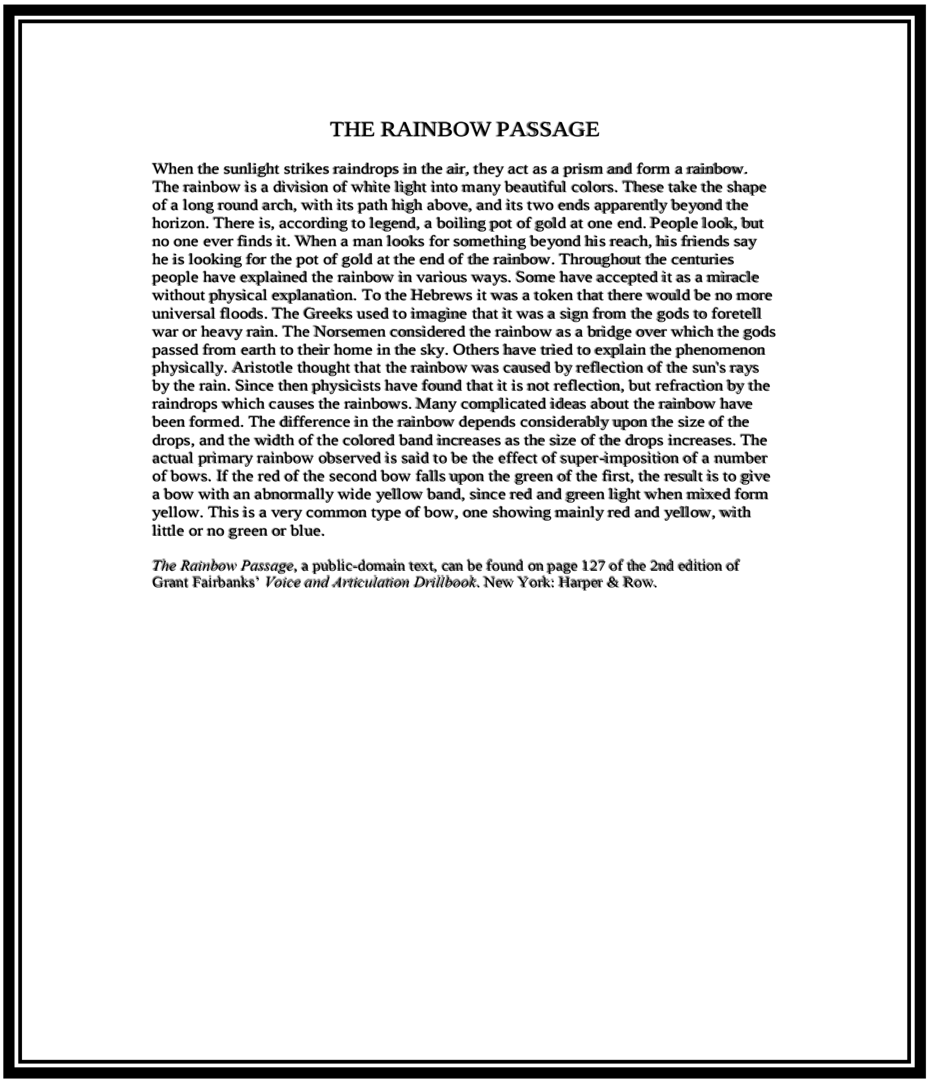
The stimuli for the pre-, mid-, and post-tests were identical. Each test included four tasks: production, identification, auditory discrimination, and category discrimination. Below are the stimuli utilised for each task, displayed in a randomised sequence:

Production test 1

Sets	Stimuli
/iː/, /ɪ/, /e/	seed
	built
	belt
/e/, /eɪ/, /aɪ/	melt
	plate
	dine
/ʌ/, /æ/, /ɑː/	sum
	clap
	harm
/ʌ/, /ɒ/, /ʊ/	dust
	fault
	bush
/ɜː/, /ɑː/, /ɔː/	worm
	carb
	sort
/ɜː/, /ɛə/, /e/	surf
	stair
	mess
/uː/, /aʊ/, /əʊ/	soup
	pound
	soap
/uː/, /ʊ/, /ʌ/	goose
	rook
	plug
/aʊ/, /uː/, /ʊ/	blouse
	tube
	chook
/aʊ/, /ɔɪ/, /əʊ/	cane
	coin
	cone
/ɜː/, /eɪ/, /ɛə/	merge
	tail
	fare
/ɔː/, /ɑː/, /əʊ/	board
	chart
	coach

Production test 2

Reading "*The Rainbow*" passage (Fairbanks, 1960)



THE RAINBOW PASSAGE

When the sunlight strikes raindrops in the air, they act as a prism and form a rainbow. The rainbow is a division of white light into many beautiful colors. These take the shape of a long round arch, with its path high above, and its two ends apparently beyond the horizon. There is, according to legend, a boiling pot of gold at one end. People look, but no one ever finds it. When a man looks for something beyond his reach, his friends say he is looking for the pot of gold at the end of the rainbow. Throughout the centuries people have explained the rainbow in various ways. Some have accepted it as a miracle without physical explanation. To the Hebrews it was a token that there would be no more universal floods. The Greeks used to imagine that it was a sign from the gods to foretell war or heavy rain. The Norsemen considered the rainbow as a bridge over which the gods passed from earth to their home in the sky. Others have tried to explain the phenomenon physically. Aristotle thought that the rainbow was caused by reflection of the sun's rays by the rain. Since then physicists have found that it is not reflection, but refraction by the raindrops which causes the rainbows. Many complicated ideas about the rainbow have been formed. The difference in the rainbow depends considerably upon the size of the drops, and the width of the colored band increases as the size of the drops increases. The actual primary rainbow observed is said to be the effect of super-imposition of a number of bows. If the red of the second bow falls upon the green of the first, the result is to give a bow with an abnormally wide yellow band, since red and green light when mixed form yellow. This is a very common type of bow, one showing mainly red and yellow, with little or no green or blue.

The Rainbow Passage, a public-domain text, can be found on page 127 of the 2nd edition of Grant Fairbanks' *Voice and Articulation Drillbook*. New York: Harper & Row.

Identification test

Vowels	Stimuli	Speakers	Set of responses
/iː/, /ɪ/, /e/	seal	male 1, SSBE	seal (sheet, teeth), sill (kid, rich), sell (pen, desk)
	sill	female 1, SSBE	seal (sheet, teeth), sill (kid, rich), sell (pen, desk)
	sell	male 1, SSBE	seal (sheet, teeth), sill (kid, rich), sell (pen, desk)
/e/, /eɪ/, /aɪ/	bet	female 1, SSBE	bet (pen, desk), bait (gate, taste), bite (night, right)
	bait	male 1, SSBE	bet (pen, desk), bait (gate, taste), bite (night, right)
	bite	female 1, SSBE	bet (pen, desk), bait (gate, taste), bite (night, right)
/ʌ/, /æ/, /ɑː/			
	hut	female 1, SSBE	hut (cut, luck), hat (cat, bad), heart (park, card)
	hat	male 1, SSBE	hut (cut, luck), hat (cat, bad), heart (park, card)
	heart	male 1, SSBE	hut (cut, luck), hat (cat, bad), heart (park, card)
/ʌ/, /ɒ/, /ʊ/	shuck	female 1, SSBE	shuck (cut, luck), shock (cost, job), shook (book, look)
	shock	male 1, SSBE	shuck (cut, luck), shock (cost, job), shook (book, look)
	shook	male 1, SSBE	shuck (cut, luck), shock (cost, job), shook (book, look)
/ɜː/, /ɑː/, /ɔː/			
	shirt	male 1, SSBE	shirt (birth, dirt), shart (park, card), short (bought, thought)
	shart	female 1, SSBE	shirt (birth, dirt), shart (park, card), short (bought, thought)
	short	male 1, SSBE	shirt (birth, dirt), shart (park, card), short (bought, thought)
/ɜː/, /ɛə/, /e/	Bert	female 1, SSBE	Bert (birth, dirt), bairt (care, wear), bet (pen, desk)
	bairt	male 1, SSBE	Bert (birth, dirt), bairt (care, wear), bet (pen, desk)
	bet	female 1, SSBE	Bert (birth, dirt), bairt (care, wear), bet (pen, desk)
/uː/, /aʊ/, /əʊ/	boot	male 1, SSBE	boot (food, choose), bout (town, shout), boat (note, wrote)
	bout	female 1, SSBE	boot (food, choose), bout (town, shout), boat (note, wrote)
	boat	male 1, SSBE	boot (food, choose), bout (town, shout), boat (note, wrote)
/uː/, /ʊ/, /ʌ/	suit	female 1, SSBE	suit (food, choose), soot (book, look), sut (cut, luck)
	soot	male 1, SSBE	suit (food, choose), soot (book, look), sut (cut, luck)
	sut	female 1, SSBE	suit (food, choose), soot (book, look), sut (cut, luck)
/aʊ/, /uː/, /ʊ/	bout	male 1, SSBE	bout (town, shout), boot (food, choose), boott (book, look)
	boot	female 1, SSBE	bout (town, shout), boot (food, choose), boott (book, look)
	boott	male 1, SSBE	bout (town, shout), boot (food, choose), boott (book, look)
/aʊ/, /ɔɪ/, /əʊ/	foul	female 1, SSBE	foul (town, shout), foil (choice, point), foal (note, wrote)
	foil	male 1, SSBE	foul (town, shout), foil (choice, point), foal (note, wrote)
	foal	female 1, SSBE	foul (town, shout), foil (choice, point), foal (note, wrote)
/ɜː/, /eɪ/, /ɛə/	dur	male 1, SSBE	dur (birth, dirt), day (gate, taste), dare (care, wear)
	day	female 1, SSBE	dur (birth, dirt), day (gate, taste), dare (care, wear)
	dare	male 1, SSBE	dur (birth, dirt), day (gate, taste), dare (care, wear)
/ɔː/, /ɑː/, /əʊ/	form	male 1, SSBE	form (bought, thought), farm (park, card), foam (note, wrote)
	farm	female 1, SSBE	form (bought, thought), farm (park, card), foam (note, wrote)

Auditory discrimination test

Vowels	Speaker	Stimuli
seal(/iː/, /e/)	male 1, SSBE	sell, sell, seal
shuck(/ʌ/, /ʊ/)	female 1, SSBE	shook, shook, shuck
bout(/uː/, /aʊ/)	male 1, SSBE	boot, boot, bout
Bert(/ɜː/, /eɪ/)	female 1, SSBE	bait, bait, Bert
foam(/ɑː/, /əʊ/)	male 1, SSBE	farm, farm, foam
foal(/ɔɪ/, /əʊ/)	female 1, SSBE	foil, foil, foal
boat(/aʊ/, /əʊ/)	male 1, SSBE	bout, bout, boat
sell(/ɪ/, /e/)	female 1, SSBE	sill, sill, sell
soot(/ʊ/, /ʌ/)	male 1, SSBE	sut, sut, soot
foil(/aʊ/, /ɔɪ/)	male 1, SSBE	foul, foul, foil
sill(/iː/, /ɪ/)	male 1, SSBE	seal, sill, seal
hat(/æ/, /ɑː/)	female 1, SSBE	heart, hat, heart
hut(/ʌ/, /æ/)	male 1, SSBE	hat, hut, hat
shock(/ʌ/, /ɒ/)	female 1, SSBE	shuck, shock, shuck
bet(/ɜː/, /e/)	male 1, SSBE	Bert, bet, Bert
suit(/uː/, /ʊ/)	female 1, SSBE	soot, suit, soot
boott(/aʊ/, /ʊ/)	male 1, SSBE	bout, boott, bout
short(/ɑː/, /ɔː/)	female 1, SSBE	shart, short, shart
bait(/eɪ/, /ɛə/)	male 1, SSBE	bairt, bait, bairt
bite(/aɪ/, /eɪ/)	male 1, SSBE	bite, bait, bait
shook(/ɒ/, /ʊ/)	male 1, SSBE	shook, shock, shock
shirt(/ɜː/, /ɑː/)	male 1, SSBE	shirt, shart, shart
bairt(/ɜː/, /ɛə/)	male 1, SSBE	bairt, Bert, Bert
bet(/ɛə/, /e/)	male 1, SSBE	bet, bairt, bairt
bout(/aʊ/, /ʊ/)	female 1, SSBE	bout, boott, boott
sut(/uː/, /ʌ/)	female 1, SSBE	sut, suit, suit
boot(/uː/, /ʊ/)	female 1, SSBE	boot, boott, boott
farm(/ɔː /, /ɑː/)	female 1, SSBE	farm, form, form

Category discrimination test

Vowels	Speaker	Stimuli
seal(/iː/, /e/)	male 1, SSBE	sell, bet, seal
sill(/ɪ/, /e/)	female 1, SSBE	bet, sell, sill
lid (/iː/, /ɪ/)	male 1, SSBE	seal, peep, lid
shirt(/ɜː/, /ɑː/)	female 1, SSBE	shart, farm, shirt
bairt(/ɛə/, /e/)	male 1, SSBE	bet, bent, bairt
soot(/uː/, /ʊ/)	female 1, SSBE	suit, boot, soot
foil(/aʊ/, /ɔɪ/)	male 1, SSBE	foul, bout, foil
farm(/ɔː/, /ɑː/)	female 1, SSBE	form, fork, farm
hut(/ʌ/, /æ/)	male 1, SSBE	hat, nan, hut
shuck(/ʊ/, /ʌ/)	female 1, SSBE	shook, wood, shuck
form(/ɑː/, /ɔː/)	male 1, SSBE	heart, form, farm
boott(/aʊ/, /ʊ/)	female 1, SSBE	bout, boott, foul
suit(/uː/, /ʌ/)	male 1, SSBE	sut, suit, touch
foal(/ɔɪ/, /əʊ/)	female 1, SSBE	foil, foal, voice
bairt(/eɪ/, /ɛə/)	male 1, SSBE	flake, bairt, bait
shuck(/ʌ/, /ɒ/)	female 1, SSBE	shock, shuck, rock
bait(/aɪ/, /eɪ/)	male 1, SSBE	kite, bait, bite
bet(/e/, /eɪ/)	female 1, SSBE	flake, bet, bait
short(/ɑː/, /ɔː/)	male 1, SSBE	heart, short, shart
boot(/aʊ/, /uː/)	female 1, SSBE	foul, boot, bout
hat(/æ/, /ɑː/)	male 1, SSBE	hat, heart, farm
shook(/ɒ/, /ʊ/)	female 1, SSBE	shook, rock, shock
bairt(/ɜː/, /ɛə/)	male 1, SSBE	bairt, Bert, nerd
Bert(/ɜː/, /e/)	female 1, SSBE	Bert, bet, bent
boat(/aʊ/, /əʊ/)	male 1, SSBE	boat, bout, foul
sut(/ʌ/, /ʊ/)	female 1, SSBE	sut, soot, wood
bout(/uː/, /aʊ/)	male 1, SSBE	bout, suit, boot
boott(/uː/, /ʊ/)	female 1, SSBE	boott, suit, boot
bait(/ɜː/, /eɪ/)	male 1, SSBE	bait, Bert, nerd
foam(/ɑː/, /əʊ/)	female 1, SSBE	foam, heart, farm

E.II First and second generalisation tests

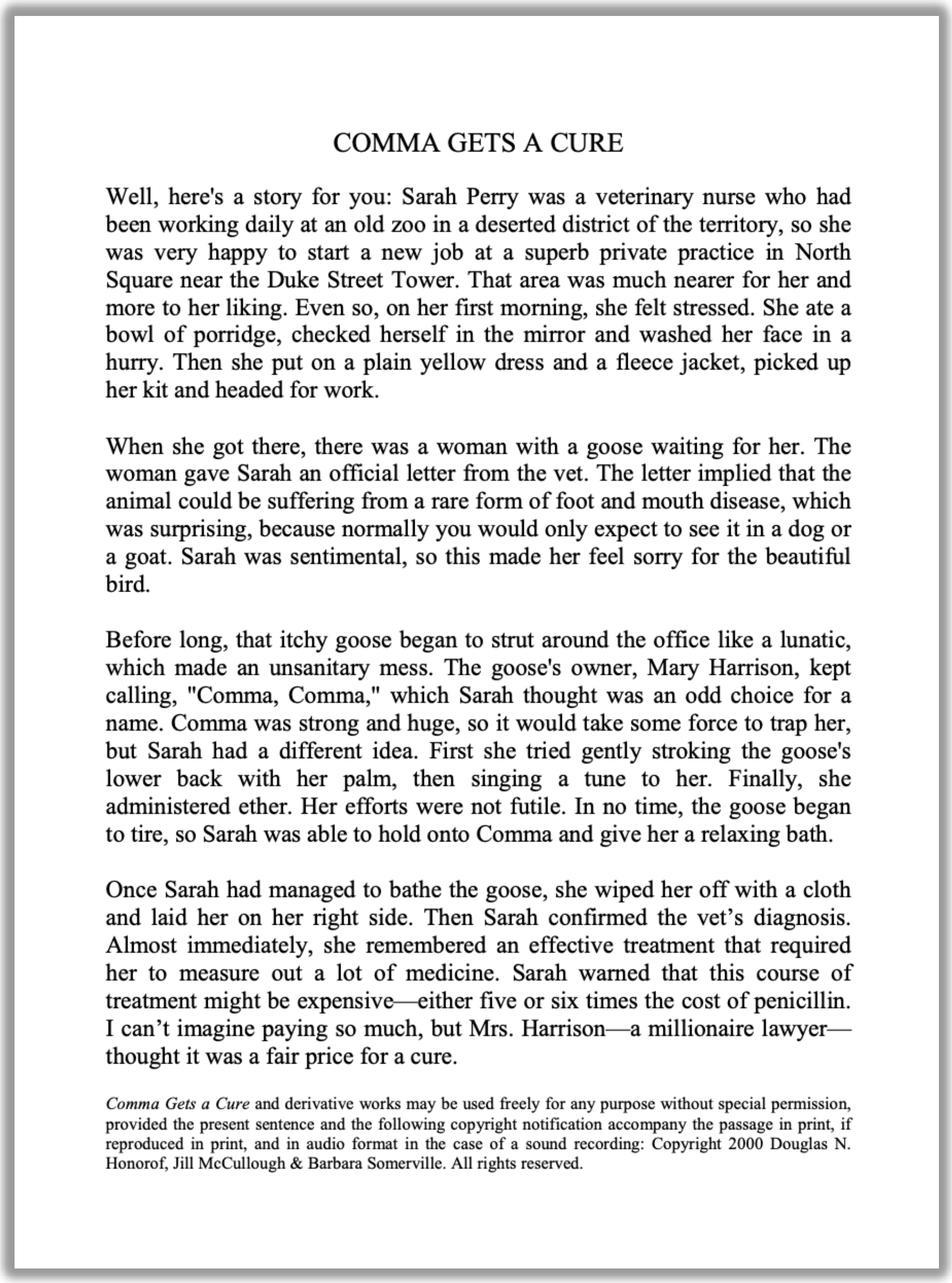
The second test of generalisation utilises the same stimuli as the first but with different accents (i.e., the words of two generalisation tests are identical but produced by different accents). Each test included four tasks: production, identification, auditory discrimination, and category discrimination. Below are the stimuli utilised for each task, displayed in a randomised sequence:

Production test 1

Sets	List of word
/iː/, /ɪ/, /e/	reach
	thin
	ted
/e/, /eɪ/, /aɪ/	vet
	great
	ripe
/ʌ/, /æ/, /ɑː/	sub
	tan
	charge
/ʌ/, /ɒ/, /ʊ/	nut
	strong
	pull
/ɜː/, /ɑː/, /ɔː/	first
	large
	stork
/ɜː/, /ɛə/, /e/	worst
	drain
	bend
/uː/, /aʊ/, /əʊ/	loop
	proud
	stone
/uː/, /ʊ/, /ʌ/	huge
	push
	gum
/aʊ/, /uː/, /ʊ/	proud
	cure
	stood
/aʊ/, /ɔɪ/, /əʊ/	mount
	spoil
	rose
/ɜː/, /eɪ/, /ɛə/	purse
	waste
	bear
/ɔː/, /ɑː/, /əʊ/	horn
	garb
	joke

Production test 2

Reading "*Comma gets a cure*” passage (Honorof et al. 2000).



COMMA GETS A CURE

Well, here's a story for you: Sarah Perry was a veterinary nurse who had been working daily at an old zoo in a deserted district of the territory, so she was very happy to start a new job at a superb private practice in North Square near the Duke Street Tower. That area was much nearer for her and more to her liking. Even so, on her first morning, she felt stressed. She ate a bowl of porridge, checked herself in the mirror and washed her face in a hurry. Then she put on a plain yellow dress and a fleece jacket, picked up her kit and headed for work.

When she got there, there was a woman with a goose waiting for her. The woman gave Sarah an official letter from the vet. The letter implied that the animal could be suffering from a rare form of foot and mouth disease, which was surprising, because normally you would only expect to see it in a dog or a goat. Sarah was sentimental, so this made her feel sorry for the beautiful bird.

Before long, that itchy goose began to strut around the office like a lunatic, which made an unsanitary mess. The goose's owner, Mary Harrison, kept calling, "Comma, Comma," which Sarah thought was an odd choice for a name. Comma was strong and huge, so it would take some force to trap her, but Sarah had a different idea. First she tried gently stroking the goose's lower back with her palm, then singing a tune to her. Finally, she administered ether. Her efforts were not futile. In no time, the goose began to tire, so Sarah was able to hold onto Comma and give her a relaxing bath.

Once Sarah had managed to bathe the goose, she wiped her off with a cloth and laid her on her right side. Then Sarah confirmed the vet’s diagnosis. Almost immediately, she remembered an effective treatment that required her to measure out a lot of medicine. Sarah warned that this course of treatment might be expensive—either five or six times the cost of penicillin. I can’t imagine paying so much, but Mrs. Harrison—a millionaire lawyer—thought it was a fair price for a cure.

Identification test for the first and second generalisation test

Vowels	List of word	Speakers	Set of responses
/iː/, /ɪ/, /e/	beat	SSBE speaker	beat (sheet, teeth), bit (kid, rich), bet (pen, desk)
	bit	Saudi speaker	beat (sheet, teeth), bit (kid, rich), bet (pen, desk)
	let	SSBE speaker	leet (sheet, teeth), lit (kid, rich), let (pen, desk)
/e/, /eɪ/, /aɪ/	stell	Saudi speaker	stell (pen, desk), stale (gate, taste), style (night, right)
	sane	SSBE speaker	sen (pen, desk), sane (gate, taste), sign (night, right)
	style	Saudi speaker	stell (pen, desk), stale (gate, taste), style (night, right)
/ʌ/, /æ/, /ɑː/			
	but	SSBE speaker	but (cut, luck), bat (cat, bad), Bart (park, card)
	hack	Saudi speaker	huck (cut, luck), hack (cat, bad), hark (park, card)
	Bart	SSBE speaker	but (cut, luck), bat (cat, bad), Bart (park, card)
/ʌ/, /ɒ/, /ʊ/	but	Saudi speaker	but (cut, luck), bot (cost, job), boott (book, look)
	bot	Saudi speaker	but (cut, luck), bot (cost, job), boott (book, look)
	hook	Saudi speaker	huck (cut, luck), hock (cost, job), hook (book, look)
/ɜː/, /ɑː/, /ɔː/			
	turn	SSBE speaker	turn (birth, dirt), tarn (park, card), torn (bought, thought)
	tarn	Saudi speaker	turn (birth, dirt), tarn (park, card), torn (bought, thought)
	torn	SSBE speaker	turn (birth, dirt), tarn (park, card), torn (bought, thought)
/ɜː/, /ɛə/, /e/	turn	Saudi speaker	turn (birth, dirt), tairn (care, wear), ten (pen, desk)
	tairn	SSBE speaker	turn (birth, dirt), tairn (care, wear), ten (pen, desk)
	ten	Saudi speaker	turn (birth, dirt), tairn (care, wear), ten (pen, desk)
/uː/, /aʊ/, /əʊ/	goot	SSBE speaker	goot (food, choose), gout (town, shout), goat (note, wrote)
	gout	Saudi speaker	goot (food, choose), gout (town, shout), goat (note, wrote)
	goat	SSBE speaker	goot (food, choose), gout (town, shout), goat (note, wrote)
/uː/, /ʊ/, /ʌ/	pool	Saudi speaker	pool (food, choose), pull (book, look), pul (cut, luck)
	pull	SSBE speaker	pool (food, choose), pull (book, look), pul (cut, luck)
	ruck	Saudi speaker	ruke (food, choose), rook (book, look), ruck (cut, luck)
/aʊ/, /uː/, /ʊ/	stoud	SSBE speaker	stoud (town, shout), stewed (food, choose), stood (book, look)
	stewed	Saudi speaker	stoud (town, shout), stewed (food, choose), stood (book, look)
	stood	SSBE speaker	stoud (town, shout), stewed (food, choose), stood (book, look)
/aʊ/, /ɔɪ/, /əʊ/	sawl	Saudi speaker	sawl (town, shout), soil (choice, point), soul (note, wrote)
	soil	SSBE speaker	sawl (town, shout), soil (choice, point), soul (note, wrote)
	soul	Saudi speaker	sawl (town, shout), soil (choice, point), soul (note, wrote)
/ɜː/, /eɪ/, /ɛə/	Kate	SSBE speaker	kate (gate, taste), curt (birth, dirt), cairt (care, wear)
	dirt	Saudi speaker	dirt (birth, bird), date (gate, taste), dairt (care, wear)
	dairt	SSBE speaker	dirt (birth, dirt), date (gate, taste), dairt (care, wear)
/ɔː/, /ɑː/, /əʊ/	port	SSBE speaker	port (bought, thought), part (park, card), poet (note, wrote)
	cark	Saudi speaker	cork (bought, thought), cark (park, card), coke (note, wrote)
	coke	SSBE speaker	cork (bought, thought), cark (park, card), coke (note, wrote)

Auditory discrimination test (1st & 2nd generalisation tests)

Vowels	Stimuli	Speakers used for the 1ST generalisation test	Speakers used for the 2nd generalisation test
lit(/ɪ/-/e/)	let, let, lit	Female 2, Saudi	Female 2, Chinese
tairn (/ɛə/, /e/)	ten, ten, tairn	Female 2, Saudi	Female 2, Chinese
gout (/uː/, /aʊ/)	goot, goot, gout	Female 2, Saudi	Female 2, Chinese
soul (/ɔɪ/, /əʊ/)	soil, soil, soul	Female 2, Saudi	Female 2, Chinese
tarn (/ɜː/, /ɑː/)	turn, turn, tarn	Female 2, Saudi	Female 2, Chinese
sign (/aɪ/- /eɪ/)	sane, sane, sign	Female 1, SSBE	Female 1, Indian
taint (/e/, /eɪ/)	tent, tent, taint	Female 1, SSBE	Female 1, Indian
stood (/aʊ/, /ʊ/)	stoud, stoud, stood	Female 1, SSBE	Female 1, Indian
bot (/ʌ/, /ɒ/)	but, but, bot	Female 2, Saudi	Female 2, Chinese
crocks (/ɒ/, /ʊ/)	crooks, crooks, crocks	Female 1, SSBE	Female 1, Indian
stoud(/aʊ/, /uː/)	stewed, stewed, stoud	Female 1, SSBE	Female 1, Indian
speed(/iː/-/ɪ/)	fit, feat, fit	Female 2, Saudi	Female 2, Chinese
ten (/ɜː/, /e/)	turn, ten, turn	Female 1, SSBE	Female 1, Indian
turn (/ɔː /, /ɜː/)	torn, turn, torn	Female 1, SSBE	Female 1, Indian
hack (/æ/, /ɑː/)	hark, hack, hark	Female 2, Saudi	Female 2, Chinese
stood (/uː/, /ʊ/)	stewed, stood, stewed	Female 2, Saudi	Female 2, Chinese
crooks (/ʊ/, /ʌ/)	crux, crooks,crux	Female 2, Saudi	Female 2, Chinese
port (/ɔː/, /ɑː/)	part, port, part	Female 1, SSBE	Female 1, Indian
soil (/aʊ/, /ɔɪ/)	sawl, soil, sawl	Female 2, Saudi	Female 2, Chinese
date (/ɜː/, /eɪ/)	dirt, date, dirt	Female 1, SSBE	Female 1, Indian
cark (/ɑː/, /əʊ/)	poem, palm, poem	Female 1, SSBE	Female 1, Indian
torn (/ɑː/, /ɔː/)	torn, tarn, tarn	Female 2, Saudi	Female 2, Chinese
turn (/ɜː/, /ɛə/)	turn, tairn, tairn	Female 1, SSBE	Female 1, Indian
pool (/uː/, /ʊ/)	pool, pull, pull	Female 2, Saudi	Female 2, Chinese
crush (/ʌ/, /æ/)	crush, crash, crash	Female 1, SSBE	Female 1, Indian
goot(/uː/, /ʌ/)	goot, gut, gut	Female 1, SSBE	Female 1, Indian
dairt(/eɪ/, /ɛə/)	dairt, date, date	Female 1, SSBE	Female 1, Indian
goat (/aʊ/, /əʊ/)	goat, gout, gout	Female 2, Saudi	Female 2, Chinese
leet(/iː/-/e/)	leet, let, let	Female 2, Saudi	Female 2, Chinese
crooks (/ʌ/, /ʊ/)	crooks, crux, crux	Female 1, SSBE	Female 1, Indian

_Category discrimination test (_1st generalisation test)

Vowels	Stimuli	Speakers
gout (/uː/, /aʊ/)	goot, pool, gout	Female 2, Saudi
ruck (/uː/, /ʌ/)	pool, sewed, ruck	Female 2, Saudi
tairn (/eɪ/, /ɛə/)	sane, taint, tairn	Female 1, SSBE
tarn (/ɑː/, /əʊ/)	goat, rode, tarn	Female 1, SSBE
crooks (/aʊ/, /ʊ/)	stoud, gout, crooks	Female 1, SSBE
wind (/ɪ/, /e/)	let, tent, wind	Female 1, SSBE
stell (/iː/, /e/)	speed, beat, stell	Female 2, Saudi
tarn (/ɔː/, /ɑː/)	torn, port, tarn	Female 1, SSBE
hack (/æ/, /ɑː/)	tarn, hark, hack	Female 2, Saudi
wine (/aɪ/, /eɪ/)	sate, wine, sane	Female 1, SSBE
jam (/e/, /æ/)	guess, jam, let	Female 1, SSBE
bot (/ʌ/, /ɒ/)	crux, bot, but	Female 2, Saudi
hook (/ʌ/, /ʊ/)	but, hook, gut	Female 1, SSBE
tairn(/ɛə/, /e/)	ten, tairn, let	Female 2, Saudi
soul (/aʊ/, /əʊ/)	gout, soul, stoud	Female 1, SSBE
port (/əʊ/, /ɔː/)	poem, port, rode	Female 1, SSBE
guess (/ɜː/, /e/)	verb, guess, herb	Female 1, SSBE
soil (/ɔɪ/, /əʊ/)	soul, soil, poem	Female 2, Saudi
beat (/iː/, /e/)	beat, let, stell	Female 2, Saudi
hook (/ɒ/, /ʊ/)	hook, hog, got	Female 1, SSBE
turn (/ɜː/, /ɛə/)	turn, tairn, dairt	Female 1, SSBE
pool (/uː/, /ʊ/)	pool, pull, hook	Female 2, Saudi
soil (/ɔɪ/, /aʊ/)	soil, sawl, gout	Female 2, Saudi
mace (/e/, /eɪ/)	mace, stell, guess	Female 1, SSBE
herb (/ɜː/, /ɛə/)	herb, dairt, tairn	Female 1, SSBE
torn (/ɔː/, /ɑː/)	torn, mars, tarn	Female 2, Saudi
goot(/uː/, /ʌ/)	goot, ruck, crush	Female 2, Saudi
stoud(/uː/, /aʊ/)	stoud, goot, pool	Female 2, Saudi

Category discrimination test (2nd generalisation test)

Vowels	Stimuli	Speakers
gout (/uː/, /aʊ/)	goot, pool, gout	Female 2, Chinese
ruck (/uː/, /ʌ/)	pool, stewed, ruck	Female 2, Chinese
pool (/uː/, /ʊ/)	hook, crooks, pool	Female 2, Chinese
tairn(/eɪ/, /ɛə/)	sane, taint, tairn	Female 1, Indian
tarn (/ɔː/, /ɑː/)	torn, port, tarn	Female 1, Indian
crooks (/aʊ/, /ʊ/)	stoud, gout, crooks	Female 1, Indian
bit (/ɪ/, /e/)	ten, let, bit	Female 1, Indian
let (/iː/, /e/)	speed, beat, let	Female 2, Chinese
tarn (/ɑː/, /əʊ/)	goat, rode, tarn	Female 1, Indian
hack (/æ/, /ɑː/)	hark, cark, hack	Female 2, Chinese
wine (/aɪ/, /eɪ/)	mace, wine, sane	Female 1, Indian
jam (/e/, /æ/)	let, jam, guess	Female 1, Indian
bot (/ʌ/, /ɒ/)	curx, bot, but	Female 2, Chinese
hook (/ʌ/, /ʊ/)	but, hook, gut	Female 1, Indian
tairn(/ɛə/, /e/)	ten, tairn, let	Female 2, Chinese
soul (/aʊ/, /əʊ/)	gout, soul, stoud	Female 1, Indian
port (/əʊ/, /ɔː/)	goat, port, rode	Female 1, Indian
guess (/ɜː/, /e/)	turn, guess, herb	Female 1, Indian
soil (/ɔɪ/, /əʊ/)	soul, soil, goat	Female 2, Chinese
beat (/iː/, /e/)	beat, tent, ten	Female 2, Chinese
pull (/ɒ/, /ʊ/)	pull, hog, got	Female 1, Indian
tarn (/ɑː/, /ɜː/)	tarn, turn, dirt	Female 2, Chinese
earth (/ɜː/, /ɛə/)	earth, dairt, tairn	Female 1, Indian
soil (/ɔɪ/, /aʊ/)	soil, sawl, gout	Female 2, Chinese
mace (/e/, /eɪ/)	mace, tent, guess	Female 1, Indian
verb (/ɜː/, /ɛə/)	verb, tairn, dairt	Female 1, Indian
torn (/ɔː/, /ɑː/)	torn, bard, hark	Female 2, Chinese
goot (/uː/, /ʌ/)	goot, crush, ruck	Female 2, Chinese
stoud (/uː/, /aʊ/)	stoud, pool, goot	Female 2, Chinese

Appendix F. Training stimuli

Stimulus Trials

Each training task was preceded by trials, which are listed in the table below. These trials were only shown in the first three sessions to aid participants in adjusting to the tasks.

Task	Stimulus Trials
Production	trim
	croak
	drawn
Auditory discrimination	greet, grit, greet
	fug, fug, fog
	pun, pan, pan
Category discrimination	maze, raise, rhyme
	lash, dark, mass
	help, street, gleam
Identification	feast (sheet, teeth), fist (kid, sick), fest (pen, desk)
	pen (pen, desk), pain (gate, taste), pine (night, right)
	feast (sheet, teeth), fist (kid, sick), fest (pen, desk)

Session 1

Production task			
Sets	**Speaker 1**	**Speaker 2**	**Speaker 3**
/iː/, /ɪ/, /e/	feel	meat	heed
	mitt	hid	fill
	head	fell	met
/e/, /eɪ/, /aɪ/	fell	met	head
	hayed	fail	mate
	might	hide	file

Auditory discrimination			
Vowel pairs	**Speaker 1**	**Speaker 2**	**Speaker 3**
/ɪ/-/e/	disk, disk, desk	mint, mint, meant	win, win, when
/iː/-/e/	den, dean, den	crepe, crepe, creep	men, mean, men
/iː/-/ɪ/	chick, cheek, cheek	deep, dip, deep	chit, cheat, cheat
/e/- /eɪ/	men, main, men	tale, tell, tell	wen, wain, wen
/aɪ/- /eɪ/	tale, tile, tale	climb, claim, claim	male, mile, male

Category discrimination			
Vowel pairs	**Speaker 1**	**Speaker 2**	**Speaker 3**
/ɪ/-/e/	miss, get, desk	peg, dill, left	hip, peg, desk
/iː/-/e/	sleep, mean, dell	keep, dean, left	get, sleep, keep
/iː/-/ɪ/	keep, mill, thick	mill, dip, beach	thick, greed, mill
/e/, /eɪ/	left, chase, tale	fade, main, get	chase, fade, peg
/aɪ/- /eɪ/	size, mice, fade	time, chase, size	tale, time, mice

Identification task		
Speaker 1		**Set of responses**
/iː/, /ɪ/, /e/	heed	heed (sheet, teeth), hid (kid, rich), head (pen, desk)
	fill	feel (sheet, teeth), fill (kid, rich), fell (pen, desk)
	met	meat (sheet, teeth), mitt (kid, rich), met (pen, desk)
/e/, /eɪ/, /aɪ/	head	head (pen, desk), hayed (gate, taste), hide (night, right)
	mate	met (pen, desk), mate (gate, taste), might (night, right)
	file	fell (pen, desk), fail (gate, taste), file (night, right)
Speaker 2		**Set of responses**
/iː/, /ɪ/, /e/	feel	feel (sheet, teeth), fill (kid, rich), fell (pen, desk)
	mitt	meat (sheet, teeth), mitt (kid, rich), met (pen, desk)
	head	heed (sheet, teeth), hid (kid, rich), head (pen, desk)
/e/, /eɪ/, /aɪ/	fell	fell (pen, desk), fail (gate, taste), file (night, right)
	hayed	head (pen, desk), hayed (gate, taste), hide (night, right)
	might	met (pen, desk), mate (gate, taste), might (night, right)
Speaker 3		**Set of responses**
/iː/, /ɪ/, /e/	meat	meat (sheet, teeth), mitt, (kid, rich), met (pen, desk)
	hid	heed (sheet, teeth), hid, (kid, rich), head (pen, desk)
	fell	feel (sheet, teeth), fill, (kid, rich), fell (pen, desk)
/e/, /eɪ/, /aɪ/	met	met (pen, desk), mate (gate, taste), might (night, right)
	fail	fell (pen, desk), fail (gate, taste), file (night, right)
	hide	head (pen, desk), hayed (gate, taste), hide (night, right)

Session 2

Production task			
Sets	**Speaker 1**	**Speaker 2**	**Speaker 3**
/ʌ/, /æ/, /ɑː/	cuck	hud	buck
	had	back	cack
	bark	shark	hard
/ʌ/, /ɒ/, /ʊ/	hud	bug	scut
	Scott	hod	lock
	look	cook	hood

Auditory discrimination			
Vowel pairs	**Speaker 1**	**Speaker 2**	**Speaker 3**
/æ/, /ɑː/	mat, mart, mat	shack, shack, shark	tarp, tap, tap
/ʌ/, /æ/	bag, bug, bug	come, cam, come	bun, bun, ban
/ʌ/, /ɒ/	sung, song, song	bonk, bonk, bunk	cup, cop, cop
/ɒ/, /ʊ/	good, good, god	should, shod, should	could, cod, could
/ʌ/, /ʊ/	cuck, cook, cuck	could, cud, cud	pus, pus, puss

Category discrimination			
Vowel pairs	**Speaker 1**	**Speaker 2**	**Speaker 3**
/æ/, /ɑː/	bank, hang, sharp	hark, shark, patch	lamb, shark, sharp
/ʌ/, /æ/	patch, gun, sank	fun, lamb, bank	sank, bank, rush
/ʌ/, /ɒ/	dog, fun, buck	son, gun, lock	fun, cod, son
/ɒ/, /ʊ/	dog, cod, crook	dog, wolf, lock	cod, brook, lock
/ʌ/, /ʊ/	crook, brook, son	sung, wolf, crook	brook, fun, wolf

Identification task		
Speaker 1		**Set of responses**
/ʌ/, /æ/, /ɑː/	buck	buck (cut, luck), back (cat, bad), bark (park, card)
	had	hud (cut, luck), had (cat, bad), hard (park, card)
	bark	buck (cut, luck), back (cat, bad), bark (park, card)
/ʌ/, /ɒ/, /ʊ/	hud	hud (cut, luck), had (cat, bad), hard (park, card)
	Scott	scut (cut, luck), Scott (cost, job), skoot (book, look)
	look	luck (cut, luck), lock (cost, job), look (book)
Speaker 2		**Set of responses**
/ʌ/, /æ/, /ɑː/	hud	hud (cut, luck), had (cat, bad), hard (park, card)
	back	buck (cut, luck), back (cat, bad), bark (park, card)
	lark	luck (cut, luck), lack (cat, bad), lark (park, card)
/ʌ/, /ɒ/, /ʊ/	luck	luck (cut), lock (cost, job), look (book, look)
	hod	hud (cut, luck), hod (cost, job), hood (book, look)
	skoot	scut (cut, luck), Scott (cost, job), skoot (book, look)
Speaker 3		**Set of responses**
/ʌ/, /æ/, /ɑː/	luck	luck (cut, luck), lack (cat, bad), lark (park, card)
	lack	luck (cut, luck), lack (cat, bad), lark (park, card)
	hard	hud (cut, luck), had (cat, bad), hard (park, card)
/ʌ/, /ɒ/, /ʊ/	scut	scut (cut, luck), Scott (cost, job), skoot (book, look)
	lock	luck (cut, luck), lock (cost, job), look (book, look)
	hood	hud (cut, luck), hod (cost, job), hood (book, look)

Session 3

Production task			
Sets	**Speaker 1**	**Speaker 2**	**Speaker 3**
/ɜː/, /ɑː/, /ɔː/	heard	burn	bird
	barn	bard	hard
	born	hoard	bored
/ɜː/, /ɛə/, /e/	burn	bird	heard
	bairn	haired	bared
	head	ben	bed

Auditory discrimination

Vowels	Speaker 1	Speaker 2	Speaker 3
/ɜː/, /ɑː/	fur, far, far	part, pert, part	barn, barn, burn
/ɑː/, /ɔː/	poor, par, poor	barn, born, born	tar, tore, tore
/ɜː/, /ɛə/	hurt, hairt, hairt	pair, purr, pair	bared, bared, bird
/ɛə/, /e/	hairt, het, hairt	bared, bed, bared	caird, ked, ked
/ɜː/, /e/	curd, curd, ked	heard, heard, head	bird, bed, bird

Category discrimination

Vowels	Speaker 1	Speaker 2	Speaker 3
/ɜː/, /ɑː/	birth, card, girl	part, nurse, curd	bard, bird, hard
/ɑː/, /ɔː/	poor, tar, par	barn, part, born	hoard, part, cart
/ɜː/, /ɛə/	fur, fair, purr	haired, girl, heard	cared, heard, pert
/ɛə/, /e/	net, dead, laird	met, haired, hell	pet, fell, hairt
/ɜː/, /e/	set, burn, beg	dead, net, birth	set, dead, heard

Identification task

Speaker 1		Set of responses
/ɜː/, /ɑː/, /ɔː/	heard	heard (birth, dirt), hard (park, card), hoard (bought, thought)
	barn	burn (birth, dirt), barn (park, card), born (bought, thought)
	born	burn (birth, dirt), barn (park, card), born (bought, thought)
/ɜː/, /ɛə/, /e/	burn	burn (birth, dirt), bairn (care, wear), ben (pen, desk)
	laird	lurd (birth, dirt), laird (care, wear), led (pen, desk)
	head	heard (birth, dirt), haired (care, wear), head (pen, desk)
Speaker 2		**Set of responses**
/ɜː/, /ɑː/, /ɔː/	burn	burn (birth, dirt), barn (park, card), born (bought, thought)
	bard	bird (birth, dirt), bard (park, card), bored (bought, thought)
	hoard	heard (birth, dirt), hard (park, card), hoard (bought, thought)
/ɜː/, /ɛə/, /e/	bird	bird (birth, dirt), bared (care, wear), bed (pen, desk)
	haired	heard (birth, dirt), haired (care, wear), head (pen, desk)
	ben	burn (birth, dirt), bairn (care, wear), ben (pen, desk)
Speaker 3		**Set of responses**
/ɜː/, /ɑː/, /ɔː/	bird	bird (birth, dirt), bard (park, card), bored (bought, thought)
	hard	heard (birth, dirt), hard (park, card), hoard (bought, thought)
	bored	bird (birth, dirt), bard (park, card), bored (bought, thought)
/ɜː/, /ɛə/, /e/	heard	heard (birth, dirt), haired (care, wear), head (pen, desk)
	bared	bird (birth, dirt), bared (care, wear), bed (pen, desk)
	bed	bird (birth, dirt), bared (care, wear), bed (pen, desk)

Session 4

Production task

Sets	Speaker 1	Speaker 2	Speaker 3
/uː/, /aʊ/, /əʊ/	who’d	roost	cooed
	cowed	how’d	roust
	code	roast	hoed
/uː/, /ʊ/, /ʌ/	cooed	poot	who’d
	put	hood	could
	hud	cud	putt

Auditory discrimination

Vowel pairs	Speaker 1	Speaker 2	Speaker 3
/uː/, /aʊ/	moose, mouse, mouse	scout, scoot, scout	who’d, how’d, how’d
/aʊ/, /əʊ/	clone, clown, clown	shout, shout, shoat	roast, roust, roust
/uː/, /ʊ/	cooed, could, cooed	who’d, who’d, hood	would, wooed, wooed
/ʊ/, /ʌ/	tuck, took, took	hud, hood, hood	putt, put, put
/uː/, /ʌ/	scut, scut, scoot	shut, shoot, shut	hud, hud, who’d

Category discrimination

Vowel pairs	Speaker 1	Speaker 2	Speaker 3
/uː/, /aʊ/	browse, bruise, bruise	cool, mouse, bruise	rule, browse, mouse
/aʊ/, /əʊ/	shout, shoat, roast	how’d, hoed, hoed	code, note, cowed
/uː/, /ʊ/	kook, cook, look	nook, rule, took	could, cooed, cook
/ʊ/, /ʌ/	mud, nook, cut	cud, put, cud	mud, dun, look
/uː/, /ʌ/	scut, scoot, who’d	cut, scoot, shoot	hud, cooed, who’d

Identification task		
Sets	**Speaker 1**	**Set of responses**
/uː/, /aʊ/, /əʊ/	who’d	who’d (food, choose), how’d (town, shout), hoed (note, wrote)
	cowed	cooed (food, choose), cowed (town, shout), code (note, wrote)
	code	cooed (food, choose), cowed (town, shout), code (note, wrote)
/uː/, /ʊ/, /ʌ/	cooed	cooed (food, choose), could (book, look), cud (cut, luck)
	put	poot (food, choose), put (book, look), putt (cut, luck)
	hud	who’d (food, choose), hood (book, look), hud (cut, luck)
	Speaker 2	**Set of responses**
/uː/, /aʊ/, /əʊ/	roost	roost (food, choose), roust (town, shout), roast (note, wrote)
	how’d	who’d (food, choose), how’d (town, shout), hoed (note, wrote)
	roast	roost (food, choose), roust (town, shout), roast (note, wrote)
/uː/, /ʊ/, /ʌ/	poot	poot (food, choose), put (book, look), putt (cut, luck)
	hood	who’d (food, choose), hood (book, look), hud (cut, luck)
	cud	cooed (food, choose), could (book, look), cud (cut, luck)
	Speaker 3	**Set of responses**
/uː/, /aʊ/, /əʊ/	cooed	cooed (food, choose), cowed (town, shout), code (note, wrote)
	roust	roost (food, choose), roust (town, shout), roast (note, wrote)
	hoed	who’d (food, choose), how’d (town, shout), hoed (note, wrote)
/uː/, /ʊ/, /ʌ/	who’d	who’d (food, choose), hood (book, look), hud (cut, luck)
	could	cooed (food, choose), could (book, look), cud (cut, luck)
	putt	poot (food, choose), put (book, look), putt (cut, luck)

Session 5

Production task			
Sets	**Speaker 1**	**Speaker 2**	**Speaker 3**
/aʊ/, /uː/, /ʊ/	how’d	pout	flout
	fool	cool	who’d
	full	full	hood
/aʊ/, /ɔɪ/, /əʊ/	cowl	how’d	loud
	hoyed	soy	coil
	sew	coal	hoed

Auditory discrimination			
Vowel pairs	**Speaker 1**	**Speaker 2**	**Speaker 3**
/aʊ/, /ɔɪ/	how’d, hoyed, hoyed	toil, toul, toul	boy, bow, boy
/ɔɪ/, /əʊ/	boil, boil, boll	noise, noise, nose	Lloyd, load, Lloyd
/aʊ/, /uː/	cowl, cool, cowl	how’d, how’d, who’d	flute, flout, flout
/uː/, /ʊ/	hood, who’d, hood	luke, look, look	floot, flute, floot
/aʊ/, /ʊ/	pout, pout, put	shout, should, shout	hood, how’d, how’d

Category discrimination			
Vowel pairs	**Speaker 1**	**Speaker 2**	**Speaker 3**
/uː/, /aʊ/	browse, bruise, bruise	cool, mouse, bruise	rule, browse, mouse
/aʊ/, /əʊ/	shout, shoat, roast	how’d, hoed, note	code, note, cowed
/uː/, /ʊ/	kook, cook, look	push, rule, took	could, cooed, cook
/ʊ/, /ʌ/	mud, push, cut	cud, put, mud	mud, cut, look
/uː/, /ʌ/	scut, scoot, who’d	cut, scoot, shoot	hud, cooed, who’d

Identification task		
Sets	**Speaker 1**	**Set of responses**
/aʊ/, /uː/, /ʊ/	how’d	how’d (town, shout), who’d (food, choose), hood (book, look)
	fool	foul (town, shout), fool (food, choose), full (book, look)
	kull	coal (note, wrote), cool (food, choose), kull (book, look)
/aʊ/, /ɔɪ/, /əʊ/	cowl	cowl (town, shout), coil (choice, point), coal (note, wrote)
	hoyed	how’d (town, shout), hoyed (choice, point), hoed (note, wrote)
	tote	tout (town, shout), toit (choice, point), tote (note, wrote)
/aʊ/, /uː/, /ʊ/	**Speaker 2**	**Set of responses**
	pout	pout (town, shout), poot (food, choose), put (book, look)
	cool	cowel (town, shout), cool (food, choose), kull (book, look)
	hood	how’d (town, shout), who’d (food, choose), hood (book, look)
/aʊ/, /ɔɪ/, /əʊ/	how’d	how’d (town, shout), hoyed (choice, point), hoed (note, wrote)
	toit	tout (town, shout), toit (choice, point), tote (note, wrote)
	coal	cowl (town, shout), coil (choice, point), coal (note, wrote)
/aʊ/, /uː/, /ʊ/	**Speaker 3**	**Set of responses**
	tout	tout (town, shout), toit (choice, point), tote (note, wrote)
	who’d	how’d (town, shout), who’d (food, choose), hood (book, look)
	kull	cowel (town, shout), cool (food, choose), kull (book, look)
/aʊ/, /ɔɪ/, /əʊ/	loud	loud (town, shout), Lloyd (choice, point), load (note, wrote)
	coil	cowl (town, shout), coil (choice, point), coal (note, wrote)
	hoed	how’d (town, shout), hoyed (choice, point), hoed (note, wrote)

Session 6

Production task

Sets	Speaker 1	Speaker 2	Speaker 3
/ɜː/, /eɪ/, /ɛə/	curt	heard	germ
	laid	fame	hayed
	haired	laird	cared
/ɔː/, /ɑː/, /əʊ/	lord	hoard	caught
	cart	lard	hard
	hoed	coat	load

Auditory discrimination

Vowel pairs	Speaker 1	Speaker 2	Speaker 3
/ɜː/, /eɪ/	learn, lane, lane	fame, firm, fame	laid, laid, lurd
/eɪ/, /ɛə/	pear, pear, pay	hayed, haired, hayed	cade, cade, cared
/ɔː/, /ɑː/	bard, bored, bard	parch, parch, porch	stalk, stark, stark
/ɑː/, /əʊ/	hoed, hard, hard	lard, load, lard	bone, barn, barn
/ɔː/, /əʊ/	coat, coat, caught	bode, bored, bode	hoed, hoed, hoard

Category discrimination

Vowel pairs	Speaker 1	Speaker 2	Speaker 3
/ɜː/, /eɪ/	germ, late, hurt	claim, skirt, nurse	skirt, skate, late
/eɪ/, /ɛə/	cade, cared, laired	pear, fair, pay	laird, haired, cade
/ɔː/, /ɑː/	start, pork, snore	stark, born, sport	snore, stark, sport
/ɑː/, /əʊ/	hoed, hard, bode	code, hoed, stark	card, code, bode
/ɔː /, /əʊ/	mode, snore, sport	road, pork, snore	snore, coal, born

Identification task		
Sets	**Speaker 1**	**Set of responses**
/ɜː/, /eɪ/, /ɛə/	curt	curt (birth, dirt), Kate (gate, taste), cairt (care, wear)
	laid	lurd (birth, dirt), laid (gate, taste), laird (care, wear)
	haired	heard (birth, dirt), hayed (gate, taste), haired (care, wear)
/ɔː/, /ɑː/, /əʊ/	lord	lord (bought, thought), lard (park, card), load (note, wrote)
	cart	caught (bought, thought), cart (park, card), coat (note, wrote)
	hoed	hoard (bought, thought), hard (park, card), hoed (note, wrote)
/ɜː/, /eɪ/, /ɛə/	**Speaker 2**	Set of responses
	heard	heard (birth, dirt), hayed (gate, taste), haired (care, wear)
	fame	firm (birth, dirt), fame (gate, taste), fairm (care, wear)
	laird	lurd (birth, dirt), laid (gate, taste), laird (care, wear)
/ɔː/, /ɑː/, /əʊ/	hoard	hoard (bought, thought), hard (park, card), hoed (note, wrote)
	lard	Lord (bought, thought), lard (park, card), load (note, wrote)
	coat	caught (bought, thought), cart (park, card), coat (note, wrote)
/ɜː/, /eɪ/, /ɛə/	**Speaker 3**	Set of responses
	lurd	lurd (birth, dirt), laid (gate, taste), laird (care, wear)
	hayed	heard (birth, dirt), hayed (gate, taste), haired (care, wear)
	cairt	curt (birth, dirt), Kate (gate, taste), cairt (care, wear)
/ɔː/, /ɑː/, /əʊ/	caught	caught (bought, thought), cart (park, card), coat (note, wrote)
	hard	hoard (bought, thought), hard (park, card), hoed (note, wrote)
	load	lord (bought, thought), lard (park, card), load (note, wrote)

Session 7 (revision)

Production task		
Sets	**Speaker**	**Stimulus**
/iː/, /ɪ/, /e/	1	heed
	2	mitt
	3	fell
/e/, /eɪ/, /aɪ/	2	hayed
	1	file
	3	met
/ʌ/, /æ/, /ɑː/	1	bark
	2	hud
	3	lack
/ʌ/, /ɒ/, /ʊ/	2	luck
	1	Scott
	3	hood
/ɜː/, /ɑː/, /ɔː/	1	barn
	2	hoard
	3	bird

/ɜː/, /ɛə/, /e/	2	haired
	1	burn
	3	bed

Auditory discrimination

Vowel pairs	Speaker	Stimuli
/iː/-/e/	1	den, dean, den
/ɪ/-/e/	2	mint, meant, mint
/iː/-/ɪ/	3	chit, cheat, cheat
/e/, /eɪ/	2	tale, tell, tell
/aɪ/- /eɪ/	1	tale, tile, tale
/æ/, /ɑː/	3	tarp, tap, tap
/ʌ/, /æ/	2	come, cam, come
/ʌ/, /ɒ/	1	sung, song, song
/ɒ/, /ʊ/	3	could, cod, could
/ʌ/, /ʊ/	1	cuck, cook, cuck
/ɜː/, /ɑː/	2	part, pert, part
/ɑː/, /ɔː/	1	poor, par, poor
/ɜː/, /ɛə/	3	bared, bared, bird
/ɛə/, /e/	1	hairt, het, hairt
/ɜː/, /e/	3	bird, bed, bird

Category discrimination

Vowel pairs	Speaker	Stimuli
/iː/-/e/	1	sleep, mean, dell
/ɪ/-/e/	2	peg, dill, left
/iː/-/ɪ/	3	thick, greed, mill
/e/, /eɪ/	2	fade, main, get
/aɪ/- /eɪ/	1	size, mice, fade
/æ/, /ɑː/	3	lamb, shark, sharp
/ʌ/, /æ/	2	fun, lamb, bank
/ʌ/, /ɒ/	1	fun, dog, buck
/ɒ/, /ʊ/	3	brook, dog, cod
/ʌ/, /ʊ/	2	sung, wolf, crook
/ɜː/, /ɑː/	2	birth, card, girl
/ɑː/, /ɔː/	1	code, barn, born
/ɜː/, /ɛə/	3	care, heard, pert
/ɛə/, /e/	1	met, haired, hell
/ɜː/, /e/	3	set, burn, beg

Identification task			
Six sets	**Speaker**	**Stimulus**	**Set of responses**
/iː/, /ɪ/, /e/	1	head	heed (sheet, teeth), heed (sheet, teeth), hid (kid, rich)
	2	mitt	meat (sheet, teeth), mitt (kid, rich), met (pen, desk)
	3	fell	feel (sheet, teeth), fill, (kid, rich), fell (pen, desk)
/e/, /eɪ/, /aɪ/	2	hayed	hayed (gate, taste), hide (night, right), head (pen, desk),
	1	file	fell (pen, desk), file (night, right), fail (gate, taste),
	3	met	met (gate, taste), might (night, right), met (pen, desk)
/ʌ/, /æ/, /ɑː/	3	lack	lack (cat, bad), luck (cut, luck), lark (park, card)
	2	hud	had (cat, bad), hud (cut, luck), hard (park, card)
	1	bark	back (cat, bad), buck (cut, luck), bark (park, card)
/ʌ/, /ɒ/, /ʊ/	2	luck	luck (cut, sun), lock (cost, job), look (book, look)
	1	Scott	scut (cut, luck), Scott (cost, job), skoot (book, look)
	3	hood	hud (cut, luck), hod (cost, job), hood (book, look)
/ɜː/, /ɑː/, /ɔː/	3	bird	bird (birth, dirt), bard (park, card), bored (bought, thought)
	1	barn	burn (birth, dirt), barn (park, card), born (bought, thought)
	2	hoard	heard (birth, dirt), hard (park, card), hoard (bought, thought)
/ɜː/, /ɛə/, /e/	2	haired	heard (birth, dirt), haired (care, wear), head (pen, desk)
	1	burn	burn (birth, dirt), bairn (care, wear), ben (pen, desk)
	3	bed	bird (birth, dirt), bared (care, wear), bed (pen, desk)

Session 8 (revision)

Production task		
sets	**Speaker**	**stimulus**
/uː/, /aʊ/, /əʊ/	1	who’d
	2	cowed
	3	roast
/uː/, /ʊ/, /ʌ/	2	who’d
	1	put
	3	cud
/aʊ/, /uː/, /ʊ/	1	how’d
	2	fool
	3	full
/aʊ/, /ɔɪ/, /əʊ/	2	how’d
	1	coil
	3	sew
/ɜː/, /eɪ/, /ɛə/	1	heard
	2	state
	3	laird

/ɔː/, /ɑː/, /əʊ/	2	hoard
	1	lard
	3	coat

Auditory discrimination		
Vowel pairs	**Speaker**	**Stimulus**
/uː/, /aʊ/	2	scout, scoot, scout
/aʊ/, /əʊ/	1	clone, clown, clown
/uː/, /ʊ/	3	cooed, cooed, could
/ʊ/, /ʌ/	2	hud, hood, hood
/uː/, /ʌ/	1	scut, scut, scoot
/aʊ/, /ɔɪ/	3	boy, bow, boy
/ɔɪ/, /əʊ/	2	noise, nose, noise
/aʊ/, /uː/	1	cowl, cool, cowl
/uː/, /ʊ/	2	luke, look, look
/aʊ/, /ʊ/	3	how’d, hood, how’d
/ɜː/, /eɪ/	1	learn, lane, lane
/eɪ/, /ɛə/	3	cade, cade, cared
/ɔː/, /ɑː/	2	parch, parch, porch
/ɑː/, /əʊ/	3	bone, barn, barn
/ɔː /, /əʊ/	1	coat, coat, caught

Category discrimination		
Vowel pairs	**Speaker**	**Stimuli**
/uː/, /aʊ/	2	cool, mouse, bruise
/aʊ/, /əʊ/	1	shoat, roast, shout
/uː/, /ʊ/	3	could, cooed, should
/ʊ/, /ʌ/	2	cud, put, dun
/uː/, /ʌ/	1	scut, scoot, who’d
/aʊ/, /ɔɪ/	3	doubt, pout, toil
/ɔɪ/, /əʊ/	2	toil, hoyed, tole
/aʊ/, /uː/	1	soon, shoot, cowl
/uː/, /ʊ/	2	hood, cool, who’d
/aʊ/, /ʊ/	3	pud, how’d, cook
/ɜː/, /eɪ/	1	germ, late, hurt
/eɪ/, /ɛə/	3	laird, haired, cade
/ɔː/, /ɑː/	2	stark, born, sport
/ɑː/, /əʊ/	3	card, code, bode
/ɔː/, /əʊ/	1	mode, snore, sport

Identification task			
Sets	**Speaker**	**Stimulus**	**Set of responses**
/uː/, /aʊ/, /əʊ/	1	who’d	who’d (food, choose), how’d (town, shout), hoed (note, wrote)
	2	cowed	cooed (food, choose), cowed (town, shout), code (note, wrote)
	3	roast	roost (food, choose), roust (town, shout), roast (note, wrote)
/uː/, /ʊ/, /ʌ/	2	who’d	who’d (food, choose), hood (book, look), hud (cut, luck)
	1	put	poot (food, choose), put (book, look), putt (cut, luck)
	3	cud	cooed (food, choose), could (book, look), cud (cut, luck)
/aʊ/, /uː/, /ʊ/	1	how’d	how’d (town, shout), who’d (food, choose), hood (book, look)
	2	fool	foul (town, shout), fool (food, choose), full (book, look)
	3	kull	coal (note, wrote), cool (food, choose), kull (book, look)
/aʊ/, /ɔɪ/, /əʊ/	2	how’d	how’d (town, shout), hoyed (choice, point), hoed (note, wrote)
	1	coil	cowl (town, shout), coil (choice, point), coal (note, wrote)
	3	sew	sow (town, shout), soy (choice, point), sew (note, wrote)
/ɜː/, /eɪ/, /ɛə/	1	heard	heard (birth, dirt), hayed (gate, taste), haired (care, wear)
	2	laid	lurd (birth, dirt), laid (gate, taste), laird (care, wear)
	3	laird	lurd (birth, dirt), laid (gate, taste), laird (care, wear)
/ɔː/, /ɑː/, /əʊ/	2	hoard	hoard (bought, thought), hard (park, card), hoed (note, wrote)
	1	lard	lord (bought, thought), lard (park, card), load (note, wrote)
	3	coat	caught (bought, thought), cart (park, card), coat (note, wrote)

Session 9

Production task			
Sets	**Speaker 1**	**Speaker 2**	**Speaker 3**
/iː/, /ɪ/, /e/	bead	heed	bead
	hid	bid	will
	bed	well	head
/e/, /eɪ/, /aɪ/	well	head	bed
	bade	whale	hayed
	hide	bide	while

Auditory discrimination			
Vowel pairs	**Speaker 1**	**Speaker 2**	**Speaker 3**
/ɪ/-/e/	heck, heck, hick	dead, did, dead	wit, wet, wet
/iː/-/e/	fell, fell, feel	neat, neat, net	breed, breed, bred
/iː/-/ɪ/	did, deed, did	deed, did, did	pill, pill, peel
/e/, /eɪ/	net, night, net	hell, hail, hell	wet, wait, wet
/aɪ/- /eɪ/	fight, fate, fate	like, lake, lake	fate, fight, fight

Category discrimination			
Vowel pairs	**Speaker 1**	**Speaker 2**	**Speaker 3**
/ɪ/-/e/	pick, felt, pill	pill, did, bless	wet, fill, pill
/iː/-/e/	fleck, scene, net	bent, keen, wet	fleek, fleck, bred
/iː/-/ɪ/	wit, pick, peel	deed, hick, did	keen, wit, hick
/e/, /eɪ/	pale, skate, bred	sake, rain, bent	sake, pale, net
/aɪ/- /eɪ/	fate, lime, night	mind, gate, fight	lime, rain, mind

Identification task		
Sets	**Speaker 1**	**Set of responses**
/iː/, /ɪ/, /e/	bead	bead (sheet, teeth), bid (kid, rich), bed (pen, desk)
	hid	heed (sheet, teeth), hid (kid, rich), head (pen, desk)
	bed	bead (sheet, teeth), bid (kid, rich), bed (pen, desk)
/e/, /eɪ/, /aɪ/	well	well (pen, desk), whale (gate, taste), while (night, right)
	bade	bed (pen, desk), bade (gate, taste), bide (night, right)
	hide	head (pen, desk), hayed (gate, taste), hide (night, right)
/iː/, /ɪ/, /e/	**Speaker 2**	**Set of responses**
	heed	heed (sheet, teeth), hid (kid, rich), head (pen, desk)
	bid	bead (sheet, teeth), bid (kid, rich), bed (pen, desk)
	well	wheel (sheet, teeth), will (kid, rich), well (pen, desk)
/e/, /eɪ/, /aɪ/	head	head (pen, desk), hayed (gate, taste), hide (night, right)
	whale	well (pen, desk), whale (gate, taste), while (night, right)
	bide	bed (pen, desk), bade (gate, taste), bide (night, right)
/iː/, /ɪ/, /e/	**Speaker 3**	**Set of responses**
	bead	bead (sheet, teeth), bid (kid, rich), bed (pen, desk)
	hid	heed (sheet, teeth), hid (kid, rich), head (pen, desk)
	head	heed (sheet, teeth), hid (kid, rich), head (pen, desk)
/e/, /eɪ/, /aɪ/	bed	bed (pen, desk), bade (gate, taste), bide (night, right)
	hayed	head (pen, desk), hayed (gate, taste), hide (night, right)
	while	well (pen, desk), whale (gate, taste), while (night, right)

Session 10

Production task			
Vowels	**Speaker 1**	**Speaker 2**	**Speaker**
/ʌ/, /æ/, /ɑː/	putt	cuck	hud
	had	pat	cack
	harsh	hard	part
/ʌ/, /ɒ/, /ʊ/	hud	cuck	putt
	pot	hod	cock
	cook	put	hood

Auditory discrimination			
Vowel pairs	**Speaker 1**	**Speaker 2**	**Speaker 3**
/æ/, /ɑː/	mash, mash, marsh	sark, sack, sack	had, hard, had
/ʌ/, /æ/	brush, brush, brash	suck, suck, sack	truck, truck, track
/ʌ/, /ɒ/	fond, fund, fond	doll, doll, dull	pup, pop, pop
/ɒ/, /ʊ/	pot, put, put	hood, hod, hood	cock, cook, cook
/ʌ/, /ʊ/	tuck, took, tuck	cud, could, cud	hood, hud, hud

Category discrimination			
Vowel pairs	**Speaker 1**	**Speaker 2**	**Speaker 3**
/ɜː/, /ɑː/	heard, curve, hard	nurse, carve, pert	dart, hurt, birth
/ɑː/, /ɔː/	hard, hoard, cart	cart, caught, tart	taught, hard, tart
/ɜː/, /ɛə/	fur, learn, purr	haired, girl, heard	fur, cared, purr
/ɛə/, /e/	led, red, caird	led, shell, haired	bairk, red, shell
/ɜː/, /e/	curd, ked, shell	pet, beck, pert	heard, dead, pet

Identification task		
Sets	**Speaker**	**Set of responses**
/ɜː/, /ɑː/, /ɔː/	curt	curt (birth, dirt), cart (park, card), court (bought, thought)
	star	stir (birth, dirt), star (park, card), store (bought, thought)
	hoard	heard (birth, dirt), hard (park, card), hoard (bought, thought)
/ɜː/, /ɛə/, /e/	berk	berk (birth, dirt), bairk (care, wear), beck (pen, desk)
	haired	heard (birth, dirt), haired (care, wear), head (pen, desk)
	pet	pert (birth, dirt), pairt (care, wear), pet (pen, desk)
/ɜː/, /ɑː/, /ɔː/	**Speaker 2**	**Set of responses**
	stir	stir (birth, dirt), star (park, card), store (bought, thought)
	hard	heard (birth, dirt), hard (park, card), hoard (bought, thought)
	court	curt (birth, dirt), cart (park, card), court (bought, thought)
/ɜː/, /ɛə/, /e/	heard	heard (birth, dirt), haired (care, wear), head (pen, desk)
	pairt	pert (birth, dirt), pairt (care, wear), pet (pen, desk)
	beck	berk (birth, dirt), bairk (care, wear), beck (pen, desk)
/ɜː/, /ɑː/, /ɔː/	**Speaker 3**	**Set of responses**
	heard	heard (birth, dirt), hard (park, card), hoard (bought, thought)
	cart	curt (birth, dirt), cart (park, card), court (bought, thought)
	store	stir (birth, dirt), star (park, card), store (bought, thought)
/ɜː/, /ɛə/, /e/	pert	pert (birth, dirt), pairt (care, wear), pet (pen, desk)
	pairt	pert (birth, dirt), pairt (care, wear), pet (pen, desk)
	head	heard (birth, dirt), haired (care, wear), head (pen, desk)

Session 11

Production task			
Sets	**Speaker 1**	**Speaker 2**	**Speaker 3**
/ɜː/, /ɑː/, /ɔː/	curt	stir	heard
	star	hard	cart
	hoard	court	store
/ɜː/, /ɛə/, /e/	berk	heard	pert
	haired	fair	chair
	pet	beck	head

Auditory discrimination			
Vowel pairs	**Speaker 1**	**Speaker 2**	**Speaker 3**
/ɜː/, /ɑː/	carve, carve, curve	pert, part, part	heard, hard, hard
/ɑː/, /ɔː/	four, far, four	caught, caught, cart	taught, tart, taught
/ɜː/, /ɛə/	chirr, chair, chair	fair, fair, fur	chirr, chair, chair
/ɛə/, /e/	ked, caird, ked	haired, head, haired	ked, caird, caird
/ɜː/, /e/	curd, curd, ked	pert, pet, pert	heard, heard, head

Category discrimination			
Vowel pairs	**Speaker 1**	**Speaker 2**	**Speaker 3**
/ɜː/, /ɑː/	heard, curve, hard	nurse, carve, pert	dart, hurt, birth
/ɑː/, /ɔː/	hard, hoard, cart	cart, caught, tart	taught, hard, tart
/ɜː/, /ɛə/	fur, learn, purr	haired, girl, heard	fur, cared, purr
/ɛə/, /e/	led, red, caird	led, shell, haired	bairk, red, shell
/ɜː/, /e/	curd, ked, shell	pet, beck, pert	heard, dead, pet

Identification task		
Sets	**Speaker**	**Set of responses**
/ɜː/, /ɑː/, /ɔː/	curt	curt (birth, dirt), cart (park, card), court (bought, thought)
	star	stir (birth, dirt), star (park, card), store (bought, thought)
	hoard	heard (birth, dirt), hard (park, card), hoard (bought, thought)
/ɜː/, /ɛə/, /e/	berk	berk (birth, dirt), bairk (care, wear), beck (pen, desk)
	haired	heard (birth, dirt), haired (care, wear), head (pen, desk)
	pet	pert (birth, dirt), pairt (care, wear), pet (pen, desk)
/ɜː/, /ɑː/, /ɔː/	**Speaker 2**	**Set of responses**
	stir	stir (birth, dirt), star (park, card), store (bought, thought)
	hard	heard (birth, dirt), hard (park, card), hoard (bought, thought)
	court	curt (birth, dirt), cart (park, card), court (bought, thought)
/ɜː/, /ɛə/, /e/	heard	heard (birth, dirt), haired (care, wear), head (pen, desk)
	pairt	pert (birth, dirt), pairt (care, wear), pet (pen, desk)
	beck	berk (birth, dirt), bairk (care, wear), beck (pen, desk)
/ɜː/, /ɑː/, /ɔː/	**Speaker 3**	**Set of responses**
	heard	heard (birth, dirt), hard (park, card), hoard (bought, thought)
	cart	curt (birth, dirt), cart (park, card), court (bought, thought)
	store	stir (birth, dirt), star (park, card), store (bought, thought)
/ɜː/, /ɛə/, /e/	pert	pert (birth, dirt), pairt (care, wear), pet (pen, desk)
	pairt	pert (birth, dirt), pairt (care, wear), pet (pen, desk)
	head	heard (birth, dirt), haired (care, wear), head (pen, desk)

Session 12

Production task			
Sets	**Speaker 1**	**Speaker 2**	**Speaker 3**
/uː/, /aʊ/, /əʊ/	who’d	root	booed
	rout	town	how’d
	bode	hoed	wrote
/uː/, /ʊ/, /ʌ/	booed	who’d	kook
	cook	foot	hood
	hud	cuck	bud

Auditory discrimination			
Vowel pairs	**Speaker 1**	**Speaker 2**	**Speaker 3**
/uː/, /aʊ/	cout, coot, coot	toon, toon, town	root, rout, root
/aʊ/, /əʊ/	now, now, know	browse, brose, browse	vote, vout, vout
/uː/, /ʊ/	wooed, would, wooed	cook, kook, kook	who’d, who’d, hood
/ʊ/, /ʌ/	good, good, gud	foot, phut, foot	hud, hood, hood
/uː/, /ʌ/	bun, boon, bun	hud, hud, who’d	mood, mud, mud

Category discrimination			
Vowel pairs	**Speaker 1**	**Speaker 2**	**Speaker 3**
/uː/, /aʊ/	rout, how’d, who’d	town, rout, root	vout, how’d, booed
/aʊ/,/əʊ/	hoed, bone, how’d	how’d, hoed, coat	town, tone, bone
/uː/, /ʊ/	should, wooed, book	book, group, put	kook, cook, put
/ʊ/, /ʌ/	pud, dun, luck	hud, cook, cut	phut, foot, dun
/uː/, /ʌ/	groom, rut, nuke	group, truth, mud	hud, groom, truth

Identification task		
Sets	**Speaker 1**	**Set of responses**
/uː/, /aʊ/, /əʊ/	who’d	who’d (food, choose), how’d (town, shout), hoed (note, wrote)
	rout	root (food, choose), rout (town, shout), wrote (note, wrote)
	bode	booed (food, choose), bowed (town, shout), bode (note, wrote)
/uː/, /ʊ/, /ʌ/	booed	booed (food, choose), bood (book, look), bud (cut, luck)
	cook	kook (food, choose), cook (book, look), cuck (cut, luck)
	hud	who’d (food, choose), hood (book, look), hud (cut, luck)
/uː/, /aʊ/, /əʊ/	**Speaker 2**	**Set of responses**
	root	root (food, choose), rout (town, shout), wrote (note, wrote)
	town	toon (food, choose), town (shout), tone (note, wrote)
	hoed	who’d (food, choose), how’d (town, shout), hoed (note, wrote)
/uː/, /ʊ/, /ʌ/	who’d	who’d (food, choose), hood (book, look), hud (cut, luck)
	bood	booed (food, choose), bood (book, look), bud (cut, luck)
	cuck	kook (food, choose), cook (book, look), cuck (cut, luck)
/uː/, /aʊ/, /əʊ/	**Speaker 3**	**Set of responses**
	booed	booed (food, choose), bowed (town, shout), bode (note, wrote)
	how’d	who’d (food, choose), how’d (town, shout), hoed (note, wrote)
	wrote	root (food, choose), rout (town, shout), wrote (note, wrote)
/uː/, /ʊ/, /ʌ/	kook	kook (food, choose), cook (book, look), cuck (cut, luck)
	hood	who’d (food, choose), hood (book, look), hud (cut, luck)
	bud	booed (food, choose), bood (book, look), bud (cut, luck)

Session 13

Production task			
Sets	**Speaker 1**	**Speaker 2**	**Speaker 3**
/aʊ/, /uː/, /ʊ/	cowed	scout	how’d
	boole	who’d	cooed
	hood	could	bull
/aʊ/, /ɔɪ/, /əʊ/	tout	how’d	scout
	hoyed	boil	toit
	bode	tote	hoed

Auditory discrimination			
Vowel pairs	**Speaker 1**	**Speaker 2**	**Speaker 3**
/aʊ/, /ɔɪ/	cow, coy, cow	bowel, bowel, boil	toil, toul, toul
/ɔɪ/, /əʊ/	toy, toy, toe	toit, toit, tote	hoed, hoyed, hoyed
/aʊ/, /uː/	crown, croon, crown	noun, noon, noun	how'd, how'd, who'd
/uː/, /ʊ/	should, shoed, should	bull, Boole, bull	pud, pood, pud
/aʊ/, /ʊ/	could, cowed, cowed	how'd, how'd, hood	put, pout, pout

Category discrimination			
Vowel pairs	**Speaker 1**	**Speaker 2**	**Speaker 3**
/aʊ/, /ɔɪ/	cowl, how'd, hoyed	house, boil, doubt	join doubt, cowl
/ɔɪ/, /əʊ/	boll, loan, toit	loan, tole, join	hoyed,hoed, tote
/aʊ/, /uː/	brown, croon, cool	fool, how'd, who'd	rude, mood, down
/uː/, /ʊ/	dude, hood, rule	cool, fool, full	bull, Boole, who'd
/aʊ/, /ʊ/	put, doubt, bull	pout, full, bull	hood, how'd, book

Identification task		
Sets	**Speaker 1**	**Set of responses**
/aʊ/, /uː/, /ʊ/	cowed	cowed (town, shout), cooed (food, choose), could (book, look)
	scoot	scout (town, shout), scoot (food, choose), skoot (book, look)
	hood	how'd (town, shout), who'd (food, choose), hood (book, look)
/aʊ/, /ɔɪ/, /əʊ/	tout	tout (town, shout), toit (choice, point), tote (note, wrote)
	hoyed	how'd (town, shout), hoyed (choice, point), hoed (note, wrote)
	coal	cowel (town, shout), coil (choice, point), coal (note, wrote)
/aʊ/, /uː/, /ʊ/	**Speaker 2**	**Set of responses**
	scout	scout (town, shout), scoot (food, choose), skoot (book, look)
	who'd	how'd (town, shout), who'd (food, choose), hood (book, look)
	could	cowed (town, shout), cooed (food, choose), could (book, look)
/aʊ/, /ɔɪ/, /əʊ/	how'd	how'd (town, shout), hoyed (choice, point), hoed (note, wrote)
	boil	bowel (town, shout), boil (choice, point), boll (note, wrote)
	tote	tout (town, shout), toit (choice, point), tote (note, wrote)
/aʊ/, /uː/, /ʊ/	**Speaker 3**	**Set of responses**
	how'd	how'd (town, shout), who'd (food, choose), hood (book, look)
	cooed	cowed (town, shout), cooed (food, choose), could (book, look)
	brose	browse (town, shout), bruise (food, choose), brose (book, look)
/aʊ/, /ɔɪ/, /əʊ/	scout	scout (town, shout), scoot (food, choose), skoot (book, look)
	toit	tout (town, shout), toit (choice, point), tote (note, wrote)
	hoed	how'd (town, shout), hoyed (choice, point), hoed (note, wrote)

Session 14

Production task			
Vowels	**Speaker 1**	**Speaker 2**	**Speaker 3**
/ɜː/, /eɪ/, /ɛə/	heard	burn	hurt
	bane	hate	hayed
	hairt	haired	bairn
/ɔː/, /ɑː/, /əʊ/	porch	hoard	cord
	hard	card	parch
	code	poach	hoed

Auditory discrimination			
Vowel pairs	**Speaker 1**	**Speaker 2**	**Speaker 3**
/ɜː/, /eɪ/	fail, fail, furl	lurd, laid, laid	heard, hayed, hayed
/eɪ/, /ɛə/	hate, hate, hairt	hayed, haired, hayed	cade, cade, cared
/ɔː/, /ɑː/	taught, tart, tart	darn, darn, dawn	hoard, hard, hard
/ɑː/, /əʊ/	hard, hoed, hard	bard, bode, bard	code, card, card
/ɔː /, /əʊ/	note, naught, note	dome, dome, dorm	hoed, hoard, hoed

Category discrimination			
Vowel pairs	**Speaker 1**	**Speaker 2**	**Speaker 3**
/ɜː/, /eɪ/	birth, curd, made	laid, lurd, furl	furl, fail, curd
/eɪ/, /ɛə/	haired, hairt, hate	haired, hayed, cared	cade, haired, cared
/ɔː/, /ɑː/	hoard, tart, cord	hoard, cord, darn	dawn, taught, hard
/ɑː/, /əʊ/	tode, hard, bode	code, barn, bode	card, code, bode
/ɔː /, /əʊ/	born, dome, hoed	born, dorm, dome	hoed, hoard, dorm

Identification task

Sets	Speaker 1	Set of responses
/ɜː/, /eɪ/, /ɛə/	heard	heard (birth, dirt), hayed (gate, taste), haired (care, wear)
	bane	burn (birth, dirt), bane (gate, taste), bairn (care, wear)
	hairt	hurt (birth, dirt), hate (gate, taste), hairt (care, wear)
/ɔː/, /ɑː/, /əʊ/	porch	porch (bought, thought), parch (park, card), poach (note, wrote)
	hard	hoard (bought, thought), hard (park, card), hoed (note, wrote)
	code	cord (bought, thought), card (park), code (note, wrote)
/ɜː/, /eɪ/, /ɛə/	**Speaker 2**	**Set of responses**
	hurt	hurt (birth, dirt), hate (gate, taste), hairt (care, wear)
	hate	hurt (birth, dirt), hate (gate, taste), hairt (care, wear)
	haired	heard (birth, dirt), hayed (gate, taste), haired (care, wear)
/ɔː/, /ɑː/, /əʊ/	hoard	hoard (bought, thought), hard (park, card), hoed (note, wrote)
	card	cord (bought, thought), card (park), code (note, wrote)
	poach	porch (bought, thought), parch (park, card), poach (note, wrote)
/ɜː/, /eɪ/, /ɛə/	**Speaker 3**	**Set of responses**
	hurt	hurt (birth, dirt), hate (gate, taste), hairt (care, wear)
	hayed	heard (birth, dirt), hayed (gate, taste), haired (care, wear)
	hairt	hurt (birth, dirt), hate (gate, taste), hairt (care, wear)
/ɔː/, /ɑː/, /əʊ/	cord	cord (bought, thought), card (park), code (note, wrote)
	parch	porch (bought, thought), parch (park, card), poach (note, wrote)
	hoed	hoard (bought, thought), hard (park, card), hoed (note, wrote)

Session 15 (revision)

Production task

Sets	Speaker	Stimulus
/iː/, /ɪ/, /e/	1	bead
	2	bid
	3	head
/e/, /eɪ/, /aɪ/	1	well
	3	hayed
	2	bide
/ʌ/, /æ/, /ɑː/	**Speaker**	**Stimulus**
	1	putt
	3	cack
	2	hard
/ʌ/, /ɒ/, /ʊ/	2	cuck
	1	pot
	3	hood
/ɜː/, /ɑː/, /ɔː/	**Speaker**	**Stimulus**
	1	curt
	2	part

	3	store
/ɜː/, /ɛə/, /e/	2	heard
	3	bairk
	1	pet

Auditory discrimination		
Vowel pairs	**Speaker**	**Stimuli**
/iː/-/e/	1	feel, fell, feel
/ɪ/-/e/	1	pick, felt, pill
/iː/-/ɪ/	3	pill, pill, peel
/e/, /eɪ/	2	hell, hail, hell
/aɪ/- /eɪ/	1	fight, fate, fate
/æ/, /ɑː/	3	pat, pat, part
/ʌ/, /æ/	2	cap, cap, cup
/ʌ/, /ɒ/	1	dock, duck, duck
/ɒ/, /ʊ/	3	cook, cook, cock
/ʌ/, /ʊ/	1	cud, could, cud
/ɜː/, /ɑː/	2	part, pert, part
/ɑː/, /ɔː/	1	four, far, four
/ɜː/, /ɛə/	3	chirr, chair, chair
/ɛə/, /e/	1	caird, ked, caird
/ɜː/, /e/	2	pert, pet, pert

Category discrimination		
Vowels	**Speaker**	**Stimuli**
/iː/-/e/	1	fleck, scene, net
/ɪ/-/e/	2	pill, did, bless
/iː/-/ɪ/	3	keen, wit, hick
/e/, /eɪ/	2	sake, rain, bent
/aɪ/- /eɪ/	1	fate, lime, fight
/æ/, /ɑː/	3	part, bad, cart
/ʌ/, /æ/	2	cam, run, cap
/ʌ/, /ɒ/	1	duck, dock, mud
/ɒ/, /ʊ/	3	cock, mod, cook
/ʌ/, /ʊ/	2	could, cud, look
/ɜː/, /ɑː/	1	carve, nurse, pert
/ɑː/, /ɔː/	2	hoard, hard, card
/ɜː/, /ɛə/	3	fur, bur, cared
/ɛə/, /e/	2	led, red, caird
/ɜː/, /e/	1	beck, dead, heard

Identification task			
Sets	**Speaker**	**Stimulus**	**Set of responses**
/iː/, /ɪ/, /e/	1	bead	bead (sheet, teeth), bid (kid, rich), bed (pen, desk)
	2	bid	bead (sheet, teeth), bid (kid, rich), bed (pen, desk)
	3	head	bead (sheet, teeth), bid (kid, rich), head (pen, desk)
/e/, /eɪ/, /aɪ/	2	bide	bed (pen, desk), bide (night, right), bade (gate, taste)
	1	well	well (pen, desk), whale (gate, taste), while (night, right)
	3	hayed	head (pen, desk), hayed (gate, taste, hide (night, right),
/ʌ/, /æ/, /ɑː/	1	putt	putt (cut, luck), pot (cost, job), put (book, look)
	2	part	putt (cut, luck), part (park, card), pat (cat, bad)
	3	had	hud (cut, luck), hard (park, card), had (cat, bad)
/ʌ/, /ɒ/, /ʊ/	2	cuck	cuck (cut, luck), cock (cost, job), cook (book, look)
	1	pot	putt (cut, luck), pot (cost, job), put (book, look)
	3	hood	hud (cut, luck), hod (cost, job), hood (book, look)
/ɜː/, /ɑː/, /ɔː/	1	curt	curt (birth, dirt), cart (park, card), court (bought, thought)
	2	hard	heard (birth, dirt), hard (park, card), hoard (bought, thought)
	3	store	stir (birth, dirt), star (park, card), store (bought, thought)
/ɜː/, /ɛə/, /e/	2	heard	heard (birth, dirt), haired (care, wear), head (pen, desk)
	3	pet	pert (birth, dirt), pairt (care, wear), pet (pen, desk)
	1	bairk	berk (birth, dirt), bairk (care, wear), beck (pen, desk)

Session 16 (revision)

Production task		
Vowels	**Speaker**	**Stimulus**
/uː/, /aʊ/, /əʊ/	1	who’d
	2	cowed
	3	wrote
/uː/, /ʊ/, /ʌ/	2	bood
	1	hud
	3	kook
/aʊ/, /uː/, /ʊ/	1	cowed
	2	who’d
	3	bull
/aʊ/, /ɔɪ/, /əʊ/	2	boil
	1	tout
	3	hoed
/ɜː/, /eɪ/, /ɛə/	1	heard

	2	hate
	3	bairn
/ɔː/, /ɑː/, /əʊ/	2	poach
	1	hard
	3	cord

Auditory discrimination		
Vowel pairs	**Speaker**	**Stimuli**
/uː/, /aʊ/	2	root, root, rout
/aʊ/, /əʊ/	1	know, now, now
/uː/, /ʊ/	3	kook, kook, cook
/ʊ/, /ʌ/	2	who’d, who’d, hood
/uː/, /ʌ/	1	good, good, gud
/aʊ/, /ɔɪ/	3	coy, cow, cow
/ɔɪ/, /əʊ/	2	toit, toit, tote
/aʊ/, /uː/	1	crown, crown, croon
/uː/, /ʊ/	2	pud, pood, pud
/aʊ/, /ʊ/	3	how’d, hood, how’d
/ɜː/, /eɪ/	1	fail, furl, fail
/eɪ/, /ɛə/	3	haired, hayed, hayed
/ɔː/, /ɑː/	2	dawn, darn, darn
/ɑː/, /əʊ/	3	code, card, card
/ɔː /, /əʊ/	1	note, note, naught

Category Discrimination		
Vowel pairs	**Speaker**	**Stimuli**
/uː/, /aʊ/	2	root, couch, rout
/aʊ/, /əʊ/	1	hoed, bone, how’d
/uː/, /ʊ/	3	cook, kook, put
/ʊ/, /ʌ/	2	hud, cut, look
/uː/, /ʌ/	1	rut, groom, nuke
/aʊ/, /ɔɪ/	3	doubt, cowed, hoyed
/ɔɪ/, /əʊ/	2	loan, tole, loin
/aʊ/, /uː/	1	brown, croon, cool
/uː/, /ʊ/	2	cool, full, fool
/aʊ/, /ʊ/	3	put, how’d, hood
/ɜː/, /eɪ/	1	birth, curd, made
/eɪ/, /ɛə/	3	cade, haired, cairt
/ɔː/, /ɑː/	2	hoard, cord, darn
/ɑː/, /əʊ/	3	code, bode, card
/ɔː /, /əʊ/	1	dote, naught, loan

Identification task

sets	Speaker	Stimulus	Set of responses
/uː/, /aʊ/, /əʊ/	1	who’d	who’d (food, choose), how’d (town, shout), hoed (note, wrote)
	2	cowed	cooed (food, choose), cowed (town, shout), code (note, wrote)
	3	wrote	root (food, choose), rout (town, shout), wrote (note, wrote)
/uː/, /ʊ/, /ʌ/	2	bood	booed (food, choose), bood (book, look), bud (cut, luck)
	1	hud	who’d (food, choose), hood (book, look), hud (cut, luck)
	3	kook	kook (food, choose), cook (book, look), cuck (cut, luck)
/aʊ/, /uː/, /ʊ/	1	how’d	how’d (town, shout), hooed (food, choose), hood (book, look)
	2	who’d	how’d (town, shout), who’d (food, choose), hood (book, look)
	3	kull	cowel (town, shout), cool (food, choose, kull (book, look)
/aʊ/, /ɔɪ/, /əʊ/	2	boil	boil (choice, point), bowel (town, shout), bole (note, wrote)
	1	tout	toit (choice, point), tout (town, shout), tote (note, wrote)
	3	hoed	hoyed (choice, point), how’d (town, shout), hoed (note, wrote)
/ɜː/, /eɪ/, /ɛə/	1	heard	heard (birth, dirt), hayed (gate, taste), haired (care, wear)
	2	hate	hurt (birth, dirt), hate (gate, taste), hairt (care, wear)
	3	bairn	burn (birth, dirt), bane (gate, taste), bairn (care, wear)
/ɔː/, /ɑː/, /əʊ/	2	poach	porch (bought, thought), parch (park, card), poach (note, wrote)
	1	hard	hoard (bought, thought), hard (park, card), hoed (note, wrote)
	3	cord	cord (bought, thought), card (park, mark), code (note, wrote)

The LingLab vowel-matching game[82]

read, rid, red
wet, wait, white

read	wet	red	rid
wet	red	wait	white
white	read	rid	wait

feed, fid, fed
set, state, sight

set	sate	fed	sate
feed	set	sight	fid
feed	fed	fid	sight

keen, pick, beg
leck, lake, like

peg	keen	leck	pick
lake	like	keen	leck
peg	like	lake	pick

[82] Note that the card dock in the actual game did not follow this sequence. The words were distributed randomly. The game was played following the completion of each training session.

dean, dill, dell
heck, pale, mind

mind	dean	heck	dill
dell	pale	dean	dell
mind	dill	heck	pale

tut, tat, tart
pus, dog, puss

puss	tut	pus	tat
tat	puss	dog	tart
tart	pus	tut	dog

snuck, snack, snark
tuck, tock, took

snuck	snark	tuck	snack
snark	tock	snuck	took
tock	took	snack	tuck

bug, mash, march

dull, doll, brook

bug	dull	doll	mash
march	doll	bug	march
brook	mash	dull	brook

duck, ham, harsh
fund, fond, cook

ham	harsh	fund	duck
fond	fund	duck	fond
cook	cook	harsh	ham

fur, far, four
hurt, hairt, het

hairt	fur	far	hairt
het	far	fur	hurt
four	het	hurt	four

purr, par, poor
curd, caird, ked

poor	purr	curd	par
ked	curd	ked	caird
poor	caird	purr	par

curve, carve, taught
chirr, chair, shell

curve	chirr	chirr	carve
taught	curve	carve	shell
taught	chair	chair	shell

burn, barn, born
birth, fair, net

barn	birth	burn	fair
barn	fair	burn	birth
born	born	net	net

coot, cout, coat
shoed, should, shut

shooed	coot	should	cout
coat	shud	coot	shut
shooed	coat	should	cout

shoot, shout, shoat
poot, put, putt

poot	shoat	put	shoot
putt	shout	shout	putt
shoat	poot	put	shoot

mood, vout, vote
wooed, would, mud

mood	wooed	vout	vote
mud	vout	wooed	would
mud	vote	would	mood

root, town, tone
groom, good, gun

town	groom	root	tone
good	gun	groom	good
gun	root	tone	town

cowed, cooed, could
sow, soy, sew

cooed	sow	cowed	soy
sow	cowed	soy	sew
cooed	could	sew	could

how’d, who’d, hood
cowl, coil, coal

hood	who’d	coil	cowl
coal	coil	how’d	who’d
hood	how’d	cowl	coal

heard, cade, caird
cord, card, code

code	cord	cade	cade
code	caired	heard	card
caird	card	cord	heard

learn, lane, lairn
dorm, darn, dome

darn	dorm	lane	lairn
darn	learn	dome	lane
learn	dorm	dome	lairn

skirt, skate, cared
stalk, stark, mode

cared	stalk	stalk	skirt
stark	skate	skirt	stark
skate	mode	mode	cared

firm, fame, haired

snore, dart, road

haired	fame	firm	road
snore	dart	snore	road
dart	fame	firm	haired

Appendix G. Interactions of tests across *vowels* for Identification (ID), Auditory Discrimination (AD), and Category Discrimination (CD) tasks

Table G.I: The interactions of tests (pre-, mid-, post-) across vowels for the ID task

Vowels (ID)	*Contrast*	*Mean Difference*	*p-value*
/ɔɪ/	pre - mid	-5%	0.000
	pre - post	-9%	0.000
	mid - post	-5%	0.000
/eɪ/	pre - mid	-5%	0.000
	pre - post	-9%	0.000
	mid - post	-5%	0.000
/aʊ/	pre - mid	-5%	0.000
	pre - post	-10%	0.000
	mid - post	-5%	0.000
/æ/	pre - mid	-5%	0.000
	pre - post	-10%	0.000
	mid - post	-5%	0.000
/ɔː/	pre - mid	-5%	0.000
	pre - post	-10%	0.000
	mid - post	-5%	0.000
/ʌ/	pre - mid	-6%	0.000
	pre - post	-11%	0.000
	mid - post	-6%	0.000
/iː/	pre - mid	-6%	0.000
	pre - post	-11%	0.000
	mid - post	-6%	0.000
/aɪ/	pre - mid	-6%	0.000
	pre - post	-12%	0.000
	mid - post	-6%	0.000
/ɑː/	pre - mid	-6%	0.000
	pre - post	-12%	0.000
	mid - post	-6%	0.000
/uː/	pre - mid	-6%	0.000
	pre - post	-12%	0.000
	mid - post	-6%	0.000
/əʊ/	pre - mid	-6%	0.000
	pre - post	-12%	0.000
	mid - post	-6%	0.000
/ʊ/	pre - mid	-6%	0.000
	pre - post	-12%	0.000
	mid - post	-6%	0.000
/ɛə/	pre - mid	-6%	0.000
	pre - post	-12%	0.000
	mid - post	-6%	0.000
/ɜː/	pre - mid	-6%	0.000

	pre - post	-12%	0.000
	mid - post	-6%	0.000
/e/	pre - mid	-6%	0.000
	pre - post	-12%	0.000
	mid - post	-6%	0.000
/ɪ/	pre - mid	-5%	0.000
	pre - post	-11%	0.000
	mid - post	-6%	0.000
/ɒ/	pre - mid	-3%	0.000
	pre - post	-7%	0.000
	mid - post	-4%	0.000

***Table G.II**: The interactions of tests (pre-, mid-, post-) across vowels for the AD task*

Vowels (AD)	*Contrast*	*Mean Difference*	*p-value*
/aɪ/	pre - post	-2%	0.000
	pre - mid	-2%	0.000
	mid- post	-0.2%	0.000
/aʊ/	pre - post	-2%	0.000
	pre - mid	-2%	0.000
	mid- post	-0.2%	0.000
/æ/	pre - post	-3%	0.000
	pre - mid	-2%	0.000
	mid- post	-0.3%	0.000
/ɑː/	pre - post	-4%	0.000
	pre - mid	-4%	0.000
	mid- post	-0.5%	0.000
/ɔː/	pre - post	-3%	0.000
	pre - mid	-3%	0.000
	mid- post	-0.4%	0.000
/ɔɪ/	pre - post	-1%	0.000
	pre - mid	-1%	0.000
	mid- post	-0.1%	0.000
/eɪ/	pre - post	-3%	0.000
	pre - mid	-3%	0.000
	mid- post	-0.3%	0.000
/iː/	pre - post	-1%	0.000
	pre - mid	-1%	0.000
	mid- post	-0.1%	0.000
/ʌ/	pre - post	-2%	0.000
	pre - mid	-2%	0.000
	mid- post	-0.2%	0.000
/ɒ/	pre - post	-3%	0.000
	pre - mid	-3%	0.000

	mid- post	-0.3%	0.000
/e/	pre - post	-7%	0.000
	pre - mid	-6%	0.000
	mid- post	-0.8%	0.000
/ɜː/	pre - post	-3%	0.000
	pre - mid	-3%	0.000
	mid- post	-0.4%	0.000
/ɛə/	pre - post	-7%	0.000
	pre - mid	-6%	0.000
	mid- post	-0.9%	0.000
/ɪ/	pre - post	-2%	0.000
	pre - mid	-2%	0.000
	mid- post	-0.2%	0.000
/ʊ/	pre - post	-3%	0.000
	pre - mid	-3%	0.000
	mid- post	-0.4%	0.000
/uː/	pre - post	-3%	0.000
	pre - mid	-2%	0.000
	mid- post	-0.3%	0.000
/əʊ/	pre - post	-2%	0.000
	pre - mid	-2%	0.000
	mid- post	-0.1%	0.000

Due to the use of raw data, the test comparisons (mid – post) revealed a high degree of statistical significance. Yet, the critical assessment lies in determining the extent of the mean difference.

***Table G.III**: The interactions of tests (pre-, mid-, post-) across vowels for the CD task*

Vowels (CD)	*Contrast*	*Mean Difference*	*p-value*
/aɪ/	pre - post	-24%	0.000
	pre - mid	-14%	0.000
	mid- post	-10%	0.000
/aʊ/	pre - post	-23%	0.000
	pre - mid	-15%	0.000
	mid- post	-9%	0.000
/æ/	pre - post	-21%	0.000
	pre - mid	-12%	0.000
	mid- post	-10%	0.000
/ɑː/	pre - post	-16%	0.000
	pre - mid	-11%	0.000
	mid- post	-5%	0.000
/ɔː/	pre - post	-24%	0.000
	pre - mid	-14%	0.000
	mid- post	-10%	0.000
/ɔɪ/	pre - post	-24%	0.000
	pre - mid	-14%	0.000
	mid- post	-10%	0.000

/eɪ/	pre - post	-24%	0.000
	pre - mid	-15%	0.000
	mid- post	-10%	0.000
/iː/	pre - post	-24%	0.000
	pre - mid	-15%	0.000
	mid- post	-9%	0.000
/ʌ/	pre - post	-24%	0.000
	pre - mid	-14%	0.000
	mid- post	-10%	0.000
/ɒ/	pre - post	-24%	0.000
	pre - mid	-15%	0.000
	mid- post	-10%	0.000
/e/	pre - post	-23%	0.000
	pre - mid	-15%	0.000
	mid- post	-9%	0.000
/ɜː/	pre - post	-21%	0.000
	pre - mid	-12%	0.000
	mid- post	-10%	0.000
/ɛə/	pre - post	-24%	0.000
	pre - mid	-14%	0.000
	mid- post	-10%	0.000
/ɪ/	pre - post	-23%	0.000
	pre - mid	-13%	0.000
	mid- post	-10%	0.000
/ʊ/	pre - post	-25%	0.000
	pre - mid	-15%	0.000
	mid- post	-10%	0.000
/uː/	pre - post	-25%	0.000
	pre - mid	-15%	0.000
	mid- post	-10%	0.000
/əʊ/	pre - post	-24%	0.000
	pre - mid	-14%	0.000
	mid- post	-10%	0.000

Table G. IV: The interactions of tests (pre-, post-, gen1) across vowels for the ID task

Vowels (ID)	*Contrast*	*Mean Difference*	*p-value*
/aɪ/	pre - post	-10%	0.000
	pre – gen1	-17%	0.000
	post- gen1	-7%	0.000
/aʊ/	pre - post	-10%	0.000
	pre – gen1	-18%	0.000
	post- gen1	-7%	0.000
/æ/	pre - post	-11%	0.000
	pre – gen1	-18%	0.000

	post- gen1	-8%	0.000
/ɑː/	pre - post	-12%	0.000
	pre – gen1	-20%	0.000
	post- gen1	-9%	0.000
/ɔː/	pre - post	-10%	0.000
	pre – gen1	-18%	0.000
	post- gen1	-7%	0.000
/ɔɪ/	pre - post	-10%	0.000
	pre – gen1	-17%	0.000
	post- gen1	-7%	0.000
/eɪ/	pre - post	-10%	0.000
	pre – gen1	-17%	0.000
	post- gen1	-7%	0.000
/iː/	pre - post	-12%	0.000
	pre – gen1	-21%	0.000
	post- gen1	-9%	0.000
/ʌ/	pre - post	-12%	0.000
	pre – gen1	-20%	0.000
	post- gen1	-9%	0.000
/ɒ/	pre - post	-9%	0.000
	pre – gen1	-19%	0.000
	post- gen1	-9%	0.000
/e/	pre - post	-12%	0.000
	pre – gen1	-22%	0.000
	post- gen1	-10%	0.000
/ɜː/	pre - post	-12%	0.000
	pre – gen1	-23%	0.000
	post- gen1	-11%	0.000
/ɛə/	pre - post	-11%	0.000
	pre – gen1	-21%	0.000
	post- gen1	-11%	0.000
/ɪ/	pre - post	-11%	0.000
	pre – gen1	-22%	0.000
	post- gen1	-11%	0.000
/ʊ/	pre - post	-12%	0.000
	pre – gen1	-23%	0.000
	post- gen1	-11%	0.000
/uː/	pre - post	-12%	0.000
	pre – gen1	-22%	0.000
	post- gen1	-10%	0.000
/əʊ/	pre - post	-22%	0.000
	pre – gen1	-22%	0.000
	post- gen1	-10%	0.000

Table G.V**: **The interactions of tests (pre-, post-, gen2) across vowels for the ID task

Vowels (ID)	*Contrast*	*Mean Difference*	*p-value*
/aɪ/	pre - post	-12%	0.000
	pre – gen2	-13%	0.000
	post- gen2	-7%	0.000
/aʊ/	pre - post	-10%	0.000
	pre – gen2	-10%	0.000
	post- gen2	-5%	0.000
/æ/	pre - post	-11%	0.000
	pre – gen2	-12%	0.000
	post- gen2	-6%	0.000
/ɑː/	pre - post	-12%	0.000
	pre – gen2	-13%	0.000
	post- gen2	-6%	0.000
/ɔː/	pre - post	-10%	0.000
	pre – gen2	-11%	0.000
	post- gen2	-5%	0.000
/ɔɪ/	pre - post	-9%	0.000
	pre – gen2	-9%	0.000
	post- gen2	-4%	0.000
/eɪ/	pre - post	-10%	0.000
	pre – gen2	-11%	0.000
	post- gen2	-5%	0.000
/iː/	pre - post	-12%	0.000
	pre – gen2	-12%	0.000
	post- gen2	-6%	0.000
/ʌ/	pre - post	-11%	0.000
	pre – gen2	-11%	0.000
	post- gen2	-5%	0.000
/ɒ/	pre - post	-9%	0.000
	pre – gen2	-9%	0.000
	post- gen2	-5%	0.000
/e/	pre - post	-12%	0.000
	pre – gen2	-13%	0.000
	post- gen2	7%	0.000
/ɜː/	pre - post	-12%	0.000
	pre – gen2	-12%	0.000
	post- gen2	7%	0.000
/ɛə/	pre - post	-11%	0.000
	pre – gen2	-12%	0.000
	post- gen2	7%	0.000
/ɪ/	pre - post	-11%	0.000
	pre – gen2	-12%	0.000
	post- gen2	7%	0.000

/ʊ/	pre - post	-12%	0.000
	pre – gen2	-13%	0.000
	post- gen2	7%	0.000
/uː/	pre - post	-12%	0.000
	pre – gen2	-13%	0.000
	post- gen2	7%	0.000
/əʊ/	pre - post	-12%	0.000
	pre – gen2	-13%	0.000
	post- gen2	-6%	0.000

Table G. VI: The interactions of tests (pre-, post-, gen1) across vowels for the AD task

Vowels (AD)	*Contrast*	*Mean Difference*	*p-value*
/aɪ/	pre - post	-2%	0.000
	pre – gen1	-2%	0.000
	post- gen1	0.1%	0.23
/aʊ/	pre - post	-2%	0.000
	pre – gen1	-1%	0.000
	post- gen1	0.1%	0.089
/æ/	pre - post	-2%	0.000
	pre – gen1	-2%	0.000
	post- gen1	0.1%	0.23
/ɑː/	pre - post	-4%	0.000
	pre – gen1	-3%	0.000
	post- gen1	0.2%	0.12
/ɔː/	pre - post	-4%	0.000
	pre – gen1	-4%	0.000
	post- gen1	0.1%	0.11
/ɔɪ/	pre - post	-2%	0.000
	pre – gen1	-2%	0.000
	post- gen1	0.1%	0.23
/eɪ/	pre - post	-2%	0.000
	pre – gen1	-2%	0.000
	post- gen1	0.1%	0.13
/iː/	pre - post	-2	0.000
	pre – gen1	-2	0.000
	post- gen1	0.1%	0.13
/ʌ/	pre - post	-4%	0.000
	pre – gen1	-4%	0.000
	post- gen1	0.2%	0.15
/ɒ/	pre - post	-2%	0.000
	pre – gen1	2%	0.000
	post- gen1	0.1%	0.12

/e/	pre - post	-6%	0.000
	pre – gen1	-6%	0.000
	post- gen1	0.3%	0.14
/ɜː/	pre - post	-3%	0.000
	pre – gen1	-3%	0.000
	post- gen1	0.1%	0.081
/ɛə/	pre - post	-4%	0.000
	pre – gen1	-4%	0.000
	post- gen1	0.2%	0.11
/ɪ/	pre - post	-2	0.000
	pre – gen1	-2	0.000
	post- gen1	0.1%	0.19
/ʊ/	pre - post	-3%	0.000
	pre – gen1	-3%	0.000
	post- gen1	0.1%	0.018
/uː/	pre - post	-3%	0.000
	pre – gen1	-3%	0.000
	post- gen1	0.1%	0.08
/əʊ/	pre - post	-2%	0.000
	pre – gen1	-2%	0.000
	post- gen1	0.2%	0.072

Note: Due to the use of raw data, the test comparisons (post –gen1) revealed a high degree of statistical significance. Yet, the critical assessment lies in determining the extent of the mean difference.

Table G.VII: The interactions of tests (pre-, post-, gen2) across vowels for the AD task

Vowels (AD)	*Contrast*	*Mean Difference*	*p-value*
/aɪ/	pre - post	-2%	0.000
	pre – gen2	-2%	0.000
	post- gen2	-0.3%	0.000
/aʊ/	pre - post	-2%	0.000
	pre – gen2	-2%	0.000
	post- gen2	-0.3%	0.000
/æ/	pre - post	-2%	0.000
	pre – gen2	-2%	0.000
	post- gen2	-0.3%	0.000
/ɑː/	pre - post	-3%	0.000
	pre – gen2	-3%	0.000
	post- gen2	-0.4%	0.000
/ɔː/	pre - post	-4%	0.000
	pre – gen2	-4%	0.000
	post- gen2	-0.5%	0.000
/ɔɪ/	pre - post	-1%	0.000
	pre – gen2	-1%	0.000

	post- gen2	-0.2%	0.000
/eɪ/	pre - post	-3%	0.000
	pre – gen2	-4%	0.000
	post- gen2	-0.4%	0.000
/iː/	pre - post	-2%	0.000
	pre – gen2	-2%	0.000
	post- gen2	-0.2%	0.000
/ʌ/	pre - post	-3	0.000
	pre – gen2	-3%	0.000
	post- gen2	-0.4%	0.000
/ɒ/	pre - post	-2%	0.000
	pre – gen2	-3%	0.000
	post- gen2	-0.3%	0.000
/e/	pre - post	-6%	0.000
	pre – gen2	-7%	0.000
	post- gen2	-0.9%	0.000
/ɜː/	pre - post	-4%	0.000
	pre – gen2	-4%	0.000
	post- gen2	-0.5%	0.000
/ɛə/	pre - post	-4%	0.000
	pre – gen2	-5%	0.000
	post- gen2	-0.6%	0.000
/ɪ/	pre - post	-2%	0.000
	pre – gen2	-2%	0.000
	post- gen2	-0.2%	0.000
/ʊ/	pre - post	-3%	0.000
	pre – gen2	-3%	0.000
	post- gen2	-0.3%	0.000
/uː/	pre - post	-4%	0.000
	pre – gen2	-4%	0.000
	post- gen2	-0.5	0.000
/əʊ/	pre - post	-2%	0.000
	pre – gen2	-2%	0.000
	post- gen2	-0.2%	0.000

Note: *Due to the use of raw data, the test comparisons (post –gen2) revealed a high degree of statistical significance. Yet, the critical assessment lies in determining the extent of the mean difference.*

***Table G.VIII**: The interactions of tests (pre-, post-, gen1) across vowels for the CD task*

Vowels (CD)	*Contrast*	*Mean Difference*	*p-value*
/aɪ/	pre - post	-24%	0.000
	pre – gen1	-18%	0.000
	post- gen1	5%	0.000
/aʊ/	pre - post	-21%	0.000
	pre – gen1	-17%	0.000
	post- gen1	4%	0.000

/æ/	pre - post	-23%	0.000
	pre – gen1	-17%	0.000
	post- gen1	5%	0.000
/ɑː/	pre - post	-17%	0.000
	pre – gen1	-14%	0.000
	post- gen1	3%	0.000
/ɔː/	pre - post	-23%	0.000
	pre – gen1	-18%	0.000
	post- gen1	5%	0.000
/ɔɪ/	pre - post	-25%	0.000
	pre – gen1	-19%	0.000
	post- gen1	5%	0.000
/eɪ/	pre - post	-24%	0.000
	pre – gen1	-18%	0.000
	post- gen1	5%	0.000
/iː/	pre - post	-23%	0.000
	pre – gen1	-19%	0.000
	post- gen1	5%	0.000
/ʌ/	pre - post	-24%	0.000
	pre – gen1	-19%	0.000
	post- gen1	5%	0.000
/ɒ/	pre - post	-25%	0.000
	pre – gen1	-19%	0.000
	post- gen1	5%	0.000
/e/	pre - post	-23%	0.000
	pre – gen1	-19%	0.000
	post- gen1	4%	0.000
/ɜː/	pre - post	-22%	0.000
	pre – gen1	-17%	0.000
	post- gen1	5%	0.000
/ɛə/	pre - post	-24%	0.000
	pre – gen1	-19%	0.000
	post- gen1	5%	0.000
/ɪ/	pre - post	-24%	0.000
	pre – gen1	-19%	0.000
	post- gen1	5%	0.000
/ʊ/	pre - post	-25%	0.000
	pre – gen1	-19%	0.000
	post- gen1	5%	0.000
/uː/	pre - post	-0.25%	0.000
	pre – gen1	-19%	0.000
	post- gen1	5%	0.000
/əʊ/	pre - post	-24%	0.000
	pre – gen1	-19%	0.000
	post- gen1	5%	0.000

***Table G. IX**: the interactions of tests (pre-, post-, gen2) across vowels for the CD task*

Vowels (CD)	*Contrast*	*Mean Difference*	*p-value*
/aɪ/	pre - post	-25%	0.000
	pre – gen2	-19%	0.000
	post- gen2	6%	0.000
/aʊ/	pre - post	-24%	0.000
	pre – gen2	-19%	0.000
	post- gen2	6%	0.000
/æ/	pre - post	-25%	0.000
	pre – gen2	-19%	0.000
	post- gen2	6%	0.000
/ɑː/	pre - post	-20%	0.000
	pre – gen2	-16%	0.000
	post- gen2	4%	0.000
/ɔː/	pre - post	-24%	0.000
	pre – gen2	-18%	0.000
	post- gen2	6%	0.000
/ɔɪ/	pre - post	-25%	0.000
	pre – gen2	-19%	0.000
	post- gen2	6%	0.000
/eɪ/	pre - post	-24%	0.000
	pre – gen2	-18%	0.000
	post- gen2	6%	0.000
/iː/	pre - post	-24%	0.000
	pre – gen2	-19%	0.000
	post- gen2	6%	0.000
/ʌ/	pre - post	-25%	0.000
	pre – gen2	-19%	0.000
	post- gen2	6%	0.000
/ɒ/	pre - post	-24%	0.000
	pre – gen2	-19%	0.000
	post- gen2	6%	0.000
/e/	pre - post	-24%	0.000
	pre – gen2	-19%	0.000
	post- gen2	6%	0.000
/ɜː/	pre - post	-23%	0.000
	pre – gen2	-17%	0.000
	post- gen2	6%	0.000
/ɛə/	pre - post	-24%	0.000
	pre – gen2	-18%	0.000
	post- gen2	6%	0.000
/ɪ/	pre - post	-24%	0.000
	pre – gen2	18%	0.000

	post- gen2	6%	0.000
/ʊ/	pre - post	-24%	0.000
	pre – gen2	-19%	0.000
	post- gen2	6%	0.000
/uː/	pre - post	-25%	0.000
	pre – gen2	-19%	0.000
	post- gen2	6%	0.000
/əʊ/	pre - post	-24%	0.000
	pre – gen2	-18%	0.000
	post- gen2	6%	0.000