December 2021

The Effect of Shadowing in Learning L2 Segments: A Perspective from Phonetic Convergence

Ruqayyah Althubyani
University of Wisconsin-Milwaukee

THE EFFECT OF SHADOWING IN LEARNING L2 SEGMENTS: A PERSPECTIVE FROM PHONETIC CONVERGENCE

by

Ruqayyah Althubyani

A Dissertation Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in Linguistics

at

The University of Wisconsin -Milwaukee

December 2021

ABSTRACT

THE EFFECT OF SHADOWING IN LEARNING L2 SEGMENTS: A PERSPECTIVE FROM PHONETIC CONVERGENCE

by

Ruqayyah Althubyani

The University of Wisconsin-Milwaukee, 2021
Under the Supervision of Professor Hanyong Park

This study aimed to investigate the role that phonetic convergence plays in the acquisition of L2 segments. In particular, it examined whether phonetic convergence towards native speakers could help Arabic-speaking second-language (L2) learners of English improve their pronunciation of four problematic English segments (/p, v, ɛ, oʊ/). To do so, the study went through several phases of experimental studies. Phonetic convergence was first explored in the productions of Arabic L2 learners towards five different English native model talkers in non-interactive setting. Five XAB perceptual similarity judgments and acoustic measurements of VOT, vowel duration, F0, and F1*F2 were used to evaluate phonetic convergence.

Based mainly on perceptual measures of phonetic convergence, learners were divided evenly between two groups. C-group (convergence group) received phonetic production training from the model talkers to whom they showed the highest degree of phonetic convergence, while D-group (divergence group) received training from the model talkers they showed divergence from or the least convergence to. Training lasted three consecutive days with target segments (i.e., /p, v, ɛ, oʊ/) presented in nonsense words. They were trained using the shadowing technique that used low-variability training paradigm in which each learner received training from one native model talker.

Native-speaker judgments on segmental intelligibility indicated both groups showed significant improvement on the post-test; however, no significant differences were found between groups in terms of the overall magnitude of this change. Perceived convergence in learners' speech failed to explain the improvement. However, some patterns of acoustic convergence towards their trainers, regardless of group, predicted the overall segmental intelligibility gains. The findings suggested that the more trainees converged their vowel duration and formants to their trainers, the more their performance improved.

At featural level, the study examined the relationship between the preexisting phonetic distance between the Arabic L2 learners of English and model talkers before the exposure and the degree of convergence. Results indicated that there was a direct relationship between how far Arabic L2 learners were from the native model talkers and the degree of convergence in all measured acoustic features. That is, the greater the baseline distance, the greater the degree of phonetic convergence was. However, such a relationship might be due to the metric used to assess phonetic convergence. The relationship between phonetic convergence measured by difference in distance (DID) and the absolute baseline distance is always biased due to the way they are calculated (Cohen Priva & Sanker, 2019; MacLeod, 2021).

This study found shadowing to be an effective technique to promote segmental intelligibility among Arabic-speakers learning English as an L2. However, this effectiveness might be increased by trainees converging more to their trainers in vowel duration and vowel spectra or being similar to their trainers in this regard from the beginning.

DEDICATION

To my parents,
my husband,
my sons,
my brothers and sisters,
and my dear grandparents' souls

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

CAT	Communication Accommodation Theory
DID	Difference-in-distance
C-group	Convergence group
D-group	Divergence group
HVPT	High-variability phonetic training
LVPT	Low-variability phonetic training
GEE	Generalized estimating equation
ICC	Intraclass correlation coefficient
M	Mean
Md	Median
SD	Standard deviation
df	Degree of freedom
K-S	Kolmogorov-Smirnov test of normality
PC	Perceived convergence

ACKNOWLEDGEMENTS

Without the support of many people, I would have never accomplished this work. I would like to express my deepest gratitude to my advisor, Hanyong Park, for all the practical and emotional guidance he gave throughout this journey. He was always meticulous, dedicated, and committed. He would answer my emails and questions any time immediately. Not only did he help me improve my academic skills and become a better researcher, he also helped me think critically and see things from different perspectives.

I would like to extend my sincere gratitude to my amazing committee members, Anne Pycha, Jae Yung Song, and Sandra Liliana Pucci, for their invaluable recommendations to make this research more successful. Many thanks also go to the faculty members in the linguistics department at the University of Wisconsin-Milwaukee who helped me grow academically and personally. I wouldn’t be where I am today without them.

This dissertation would not have been possible without the support, encouragement, and love of my husband, Salman Albardi, who was also busy with his PhD. With two children, it was challenging to balance parenthood and academics. My oldest son, Abdulrahman, was only 8 when the pandemic started but was endlessly understanding and helped take care of his younger brother, Mohammad.

Words cannot express my heartfelt appreciation to my parents and siblings for their unconditional love. They were the first to believe in me and encourage me to pursue my graduate studies. During my doctoral studies, I lost one of my sisters and my grandparents, and I dedicate this dissertation to their memory, and I wish they were here to celebrate my achievement. Many thanks also go to my mother-in-law and siblings-in-law for their support, motivation, and prayers.

I would like to thank my dear friends Amal El Haimeur and Saja Albuarabi for their friendship and love. A special thanks to Amal's mother, who was a second mother to me. They are always there for me during the most happy moments as well as the hardest time. They brightened my time in the US and helped me face hardships and obstacles.

I would also like to thank Amal Khalfaoui, one of my professors at Florida Atlantic University during my graduate studies. She encouraged me to pursue a PhD, and I learned a lot from her as a professor and a friend. I always enjoyed her company and grew from her advice.

A special thanks to my Saudi participants, as this dissertation would not have been possible without them. Their contribution is especially noteworthy since they volunteered during a pandemic and came to the linguistics department to be recorded before a vaccine was available. It was difficult to find enough participants due to the pandemic, but this task was made easier by my friends. I am deeply indebted to all my friends in Saudi Arabia and the US for their support and encouragement throughout the years.

CHAPTER 1: INTRODUCTION

1.1 Introduction

Good pronunciation of a second language (L2) plays an essential role in successful communication with native speakers, while poor pronunciation can make it difficult for non-native speakers to be understood (Celce-Murcia et al., 1996). Most L2 learners have the desire to attain more native-like pronunciation to successfully communicate with native speakers. However, this is a complicated process with many factors, such as learner age, aptitude, attitude, motivation, amount of L2 exposure, teaching methods and quality, and the effect of the first language (L1) phonology.

Given the importance of pronunciation, researchers and educators have explored different techniques to teach it. One is shadowing, which was the technique employed in the present study to train L2 learners in pronunciation. Although shadowing has been used widely in classrooms across China, Japan, and South Korea (Rost & Wilson, 2013), L2 researchers have only started giving more attention to it in recent years (Foote & McDonough, 2017).

1.2 Shadowing

Lambert (1992) defines shadowing in L1 contexts as "a paced, auditory tracking task which involves the immediate vocalization of auditorily presented stimuli, i.e., word-for-word repetition, in the same language, parrot-style, of a message presented through headphones" (p. 266). In other words, shadowing is an imitation activity where learners repeat words, phrases, or sentences as quickly as possible after hearing them (Foote, 2015; Foote & McDonough, 2017; Hamada, 2017; Hsieh et al., 2013; Luo et al., 2010; Mori, 2011; Nye & Fowler, 2003).

Previous studies have reported shadowing as an effective way to improve aspects of L2 pronunciation and other skills, such as listening and reading (e.g., Bovee & Stewart, 2009; Foote & McDonough, 2017; Hamada, 2016; Hsieh et al., 2013; Horiyama, 2012; Kadota, 2019; Wang, 2017). Nonetheless, speech shadowing appeared in other domains before being extended to language learning. Speech shadowing tasks were first reported in cognitive psychology research on selective auditory attention, particularly by Cherry (1953). Two decades later, researchers started to employ speech shadowing to examine schizophrenic deficit (e.g., Lerner, 1974). Shadowing has also been used to train simultaneous interpreters (e.g., Lambert, 1992) and in speech pathology to treat stuttering (e.g., Armson & Kalinowski, 1994; Saltuklaroglu & Kalinowski, 2011). Table 1 summarizes different types of shadowing that have been used in L1 and L2 settings, as adapted from Hamada (2017, p. 5).

Table 1. *Types of Shadowing*

Name	Procedure
Complete shadowing	Learners shadow everything speakers say.
Selective shadowing	Learners select only certain words and phrases to shadow.
Parallel reading	Learners shadow while reading a text.
Content shadowing	Learners concentrate on both shadowing and the meaning.
Mumbling	Learners silently shadow the incoming sounds without text.
Interactive shadowing	Selective shadowing and adding questions and comments to make it more natural and get more learner involvement.
Conversational shadowing	Learners repeat conversation partner's words.
Phrase shadowing	Learners shadow phrase by phrase with a slight delay.
Phonemic shadowing	Learners shadow each sound as soon as they hear it.

Complete and phonemic shadowing both mean shadowing everything that was heard (Murphey, 2001). Mumbling is similar to complete and phonemic shadowing, but learners shadow what they hear silently. It is usually used in classrooms with a large number of learners

(Hamada, 2017). As the name suggests, selective shadowing includes only particular words or phrases from what is heard (Murphey, 2001).

In parallel reading, learners repeat what they hear while looking at a script, while content shadowing refers to learners shadowing what they hear while simultaneously focusing on the meaning (Hamada, 2017). Interactive and conversational shadowing include shadowing during an interaction. While conversing, listeners shadow selected phrases or sentences from the speaker and add questions to them (Murphey, 2001). Phrase shadowing and phonemic shadowing include shadowing phrases and sounds, respectively.

A survey of literature on language learning revealed no consensus on how shadowing is different from other imitation techniques, such as repetition, mirroring, and tracking, with these terms used somewhat interchangeably in some studies despite differences between them. For example, Foote and McDonough (2017) considered shadowing the same as mirroring and tracking, while Nye and Fowler (2003) considered shadowing and imitation two distinct tasks. According to them, participants in a typical shadowing task are instructed to repeat what they hear as fast as possible, which affects the shadowed speech. Regional dialects and personal vocal habits affect the phonetic and prosodic structure of shadowed speech as well. However, in imitation tasks where participants are explicitly asked to *imitate* the speech they hear, they tend to (consciously or unconsciously) inhibit their personal speech habits. Thus, the resulting imitated speech has more phonetic and non-phonetic features of the speech being imitated.

According to Hamada (2017), shadowing and repetition have some features in common as they both involve repeating after a native speaker without looking at a script; the difference lies in the time gap. Participants repeat after a native speaker as quickly as possible in shadowing while there is some delay in repetition. Hsieh et al. (2013) indicated that shadowing also differs from repetition in terms of memory load. Repetition involves more short-term memory as learners focus on memorizing pronunciation at the word or sentence level, which might hinder their concentration during the subsequent production.

Celce-Murcia et al. (1996) differentiated mirroring, tracking, and shadowing. In mirroring, L2 learners repeat simultaneously what a native speaker says, in person or on a screen, in addition to native speaker movements, such as eye movement, gestures, and posture. However, Acton (1984) viewed mirroring as imitating only non-verbal actions, such as facial expressions, gestures, posture, and body movements, adding that mirroring should be used only after learners become more experienced in tracking. Tracking refers to learners repeating instantly what a native speaker says word by word, while native speakers' movements and gestures are not imitated (Acton, 1984; Celce-Murcia et al., 1996; Danchenko, 2011). This definition is similar to shadowing, as defined by most researchers in the field. However, Martinsen et al. (2017) described tracking as an "imitation activity in which learners listen to recordings of native speakers, follow along with subtitles or transcripts, and attempt to produce what the speakers are saying as closely as possible with as little delay as possible" (p. 665). Thus, their description of tracking is similar to that of parallel reading, in which learners repeat what a native speaker says while looking at a script.

Although most researchers define shadowing as a simultaneous repetition of what is heard, Celce-Murcia et al. (1996) considered it to be a repetition that comes slightly after the speaker, which is analogous to the description of repetition by other researchers (e.g., Hamada, 2017; Hsieh at el., 2013). According to them, the only difference between shadowing and tracking is the time lag between listening and repeating, which differs from what other researchers have asserted. However, Danchenko (2011) referred to shadowing as learners repeating simultaneously what a native speaker says while reading from a script. This description is similar to tracking activities as described by other researchers such as Martinsen et al. (2017)

Hamada (2017) also differentiated between shadowing and elicited imitation, arguing that elicited imitation can be done in one of two ways: (1) learners are instructed to repeat what they hear as precisely as possible or (2) learners are presented with an ungrammatical sentence and must repeat it using the correct structure. Shadowing can be used to help learners improve their listening skills, while Hamada defined elicited imitation as being used to assess their linguistic knowledge. Another critical difference is that shadowing focuses on phonological forms, whereas elicited imitation centers on meaning.

As demonstrated above, shadowing has been referred to in different ways and has been confused with different imitation techniques. The present study adopted a definition of shadowing as described in the majority of studies: a simultaneous repetition without looking at a script.

1.3 What Is the Ideal Voice for Learners to Imitate?

Many approaches and techniques have been proposed to help L2 learners acquire a more native-like accent or more intelligible pronunciation. However, an outstanding question has always been what voice or pronunciation L2 learners should ideally imitate. The best voices to teach pronunciation have been called “golden speakers” (Ding et al., 2019; Probst et al., 2002). From the perspective of speech technology, computer-assisted language learning systems, such as the FLUENCY system (Probst at el., 2002) and Golden Speaker Builder (GSB) (Ding et al., 2019), can help find voices that L2 learners should imitate to improve their comprehensibility and fluency.

Probst at el. (2002) studied the effect of matching learners with different native speakers using the FLUENCY system. Non-native speakers of English were asked to use FLUENCY to practice how the auxiliary verb is stressed in a sentence. The learners had intermediate proficiency with various native language backgrounds and were randomly assigned to three groups. In Group 1, learners listened to six native speakers saying five sentences each. Every learner in this group was free to choose the native speaker they preferred to imitate. In Group 2, participants were assigned to native speakers based on the similarities between native speaker learner speech rate. The matching criteria were gender and similarity in fundamental frequency (F0) and speech rate (syllables). F0 and speech rate were obtained from the learners' production of the sentence "Yes, I did want to rent a car" in their native languages. In Group 3, learners were assigned to those native speakers who had the opposite gender with the most dissimilar voices in terms of F0 and speech rate. Changing the native speaker was not allowed during training, and every learner was allowed to imitate only one native speaker. Group 2 showed the best performance. Although Group 3 practiced the sentences more often than Group 2, their performance was worse. The performance of Group 1 was the lowest even though they chose their tutors themselves. These results suggested that matching participants to native speakers with the most similar speech rate could lead to more improvement.

Probst at el. (2002) measured this improvement based on the number of syllables and phones per sentence that were pronounced correctly. The pre- and post-test scores obtained from the system were compared. Neither acoustic measurements nor perceptual judgments by native speakers were used to measure how accurately their pronunciation had improved. Furthermore, F0 and speech rate were obtained from the learners' production of the sentence *"Yes, I did want to rent a car"* in their native languages, which were not controlled. Their native languages were

Russian, German, French, Spanish, Mandarin Chinese, Arabic, Thai, and Bahasa Indonesian. The authors argued that when learners produced the sentence in their native languages, they would produce the most natural speech. The study did not have control over the segments and syllables since the sentences were read in various languages. Therefore, those criteria lack accuracy to match the learners with native speakers.

Another important point is that the trainees in Probst at el. (2002) practiced only how the auxiliary verb was stressed in sentences and there was no acoustic measurement. The FLUENCY system gave users a duration correction by informing them which syllables had a longer or shorter duration than native speakers. Improvement was measured based on a reduction of duration and phone errors, but the system identified all duration errors in the sentence, not only in auxiliary verbs. Thus, how the duration of stressed auxiliary verbs improved was unclear. Although F0, intensity, and duration are primary acoustic features for stressed syllables (Ladefoged & Johnson, 2014), Probst at el. (2002) did not examine them. Thus, the study's results are limited to how stressed the auxiliary verb was in a sentence, and it is unclear how having similar features—such as speech rate, F0, and gender—would influence the efficacy of training on segmental and suprasegmental production. It is likewise unclear what other acoustic features or factors might affect the outcomes of production training. Therefore, the present study aimed to examine phonetic convergence towards native speakers in non-native settings as one of several factors that might impact production training.

1.4 Phonetic Convergence

Speech in everyday conversation is characterized by acoustic-phonetic variation. When people talk to each other, they are exposed to these phonetic details and often adjust their speech to the phonetic details of their interlocutors. This adaptation has been investigated under different names, including phonetic convergence (e.g., Abrego-Collier et al., 2011; Kim, 2012; Kim et al., 2011; Mukherjee et al., 2019; Pardo, 2006, 2010; Pardo et al., 2013; Pardo et al., 2012), phonetic accommodation (e.g., Babel & Bulatov, 2012; Babel & McGuire, 2015; Babel et al., 2014; Foster & Cole 2020; Manker, 2016; Pouplier et al., 2014; Tobin, 2013), alignment (e.g., Costa et al., 2008; Garrod & Pickering, 2007; Mitterer, 2011; Nenkova et al., 2008; Ostrand & Chodroff 2021; Reitter & Moore, 2014), phonetic imitation (e.g., Babel, 2012; MacLeod 2021; Markham, 1997; Nielsen, 2011), phonetic adaptation (Hwang et al., 2015; Ullas et al. 2020), and audience design (e.g., Bell, 1984; Clark & Murphy, 1982; Horton & Gerrig, 2005; Mason, 2000). Throughout this paper, I use the term "phonetic convergence" for this phenomenon.

1.4.1 Theoretical Accounts of Phonetic Convergence

Researchers have used different models to account for phonetic convergence and explain the motivations and mechanisms that lead to convergence or divergence. Some have used speech perception models. For example, Pardo at el. (2012) referred to the Motor Theory by Liberman and Mattingly (1985) to explain phonetic convergence. Under this theory, there is a strong link between perception and production. This indirectly suggests that when hearing another talker, in a real conversation or even passively without oral communication, listeners perceive the talker's phonetic details, which then impact their speech production later. Although this can overtly result in imitation or phonetic convergence to that talker (Pardo at el., 2013), such models cannot account for divergence.

Other researchers have tried to explain phonetic convergence by proposing theories of social interaction and cognitive systems. One of the psycholinguistic models is the Automatic Interactive Alignment Account of Dialogue (Pickering & Garrod, 2004). In this model, the assumption is that convergence in dialogue occurs in the linguistic representations that speakers use automatically and subconsciously. However, a considerable body of research contradicts the claims of automatic alignment, and evidence has shown that automatic alignment does not happen in all interactions. For instance, individuals do not consistently show alignment to certain types of talkers. Kim et al. (2011) found that the degree of convergence in a conversation was affected by the language distance between interlocutors. Phonetic convergence was found when native speakers interacted with interlocutors who had the same L1 dialect, and more divergence was found when they interacted with native speakers with a different dialect or with non-native interlocutors. Babel (2012) revealed phonetic selectivity in the investigated vowels, and the degree of convergence was also influenced by self-rated attractiveness of the interlocutor. That is, not all vowels showed the same degree of convergence; low vowels were imitated more than high vowels when female participants rated the attractiveness of White model talkers higher. More details about this study are provided when discussing the factors that affect phonetic convergence. Babel (2012) argued that phonetic convergence is not an automatic process that happens all the time but is automatic in the sense that it occurs subconsciously.

Another cognitive theory to predict phonetic convergence is Exemplar Theory (Johnson, 1997, 2006; Pierrehumbert, 2001, 2003). Similar to Automatic Interactive Alignment, this theory implies that phonetic convergence is automatic when processing incoming speech. The central claim is that the phonetic details of each instance of a word are stored in the listener's memory. When listeners hear a similar word spoken, episodic traces of that word are activated in the memory. The average of all activated words, including recently heard ones, appear in the production. Thus, familiar voices or high-frequency words with more episodic traces activate more traces, resulting in less convergence. However, low-frequency words and unfamiliar voices activate fewer traces, which result in more phonetic convergence in subsequent production (Goldinger, 1998; Pardo et al., 2017). Nevertheless, this model does not consider social factors that affect phonetic convergence, nor does it explain the cases when phonetic divergence occurs.

The socio-psychological Communication Accommodation Theory (CAT) describes how social settings mediate speech convergence (Giles et al., 1991; Shepard et al., 2001). In this model, speech modification is considered a tool that a talker uses to adjust the social distance from interlocutors. When talkers have the desire to be socially closer to another talker, they converge to that talker. They show divergence from interlocutors, or maintain their regular speech patterns, to distinguish themselves from that other group of speakers. Generally, CAT suggests that speakers or groups' (typically unconscious) need to be socially integrated or identify themselves with others is reflected by speech convergence. In this model, speakers do not necessarily show convergence at all obtainable levels. Hence, speech accommodation in CAT primarily depends on the social motivations of the talkers. This differs from other psycholinguistic models since it views convergence as a choice speakers make to manipulate the social distance between them and their interlocutors.

1.4.2 Factors That Affect Phonetic Convergence

Numerous linguistic, situational, psychological, and social factors affect phonetic convergence. A linguistic factor is the language distance between interlocutors. For instance, Kim et al. (2011) found that close language distance (L1 and dialect match) between interlocutors facilitated phonetic convergence. Another linguistic factor is word frequency, as indicated by Goldinger (1998), who found that low-frequency words resulted in more phonetic convergence than high-frequency words. Based on an episodic model of speech perception and production, he argued that low-frequency words evoked more phonetic convergence in a shadowing task because they stimulated fewer exemplars in the speakers' lexicon, as speakers heard those words fewer times. Therefore, it should be easier for speakers to retrieve the phonetic details of low-frequency words, and because of this, phonetic convergence is more likely. Therefore, the present study used low-frequency words to evoke convergence.

A social and psychological factor in phonetic convergence is a close relationship between interlocutors. For instance, Pardo et al. (2012) found that a close relationship between roommates facilitated phonetic convergence. Additionally, likeness and attractiveness have been correlated with the degree of phonetic convergence in L1 settings. Babel (2010) indicated that when participants held positive attitudes towards the model talker, they phonetically converged more to that talker. Babel's (2009, 2012) findings showed phonetic selectivity in the investigated vowels, and the degree of convergence was also influenced by self-rated attractiveness of the interlocutor. Phonetic selectivity here means that not all vowels showed the same degree of convergence, as low vowels were imitated more than high vowels. In the experimental condition where a digital image of the model talker was presented, participants were asked to rate the attractiveness of the model talker. Two digital images were used: one for a Black model talker, and one for a White model talker. In the case of the White model talker, the higher the female participants rated his attractiveness, the more convergence was found in their vowels. Male participants showed the opposite pattern: the higher the attractiveness rating, the lower convergence was found. However, there was no significant correlation for the Black model talker.

Pardo (2006) showed that phonetic convergence could be affected by gender and specific conversational roles, such as being the receiver or giver of instructions in a map task. In general, the study indicated that female speakers showed less convergence than males. Female givers showed more convergence to female receivers, but female receivers did not show convergence to female givers. In contrast, in the male pairs, male receivers exhibited convergence to givers, and male givers converged to receivers, but to a less extent.

Another factor affecting degree of convergence is any preexisting phonetic distance between baseline productions and the model talker's productions in a certain acoustic dimension (Kim, 2012). The baseline productions are produced by the participants before exposure to the model talkers. Preexisting phonetic distance can be calculated by subtracting the value of the baseline productions from the model talkers' value (i.e., the model talker's value minus the baseline value). The resulting absolute value indicates how far the baseline is from the model talker's value in a certain acoustic dimension. Kim (2012) examined phonetic convergence patterns that native English speakers exhibited towards model talkers with three linguistic distances. The participants were exposed in a non-interactive setting to a native English model talker with the same L1 dialect, a native English model talker with the same L1 but a different dialect, and a high proficiency non-native speaker. Participants were asked to read monosyllabic words, disyllabic words, and full sentences before and after exposure to the model talkers.

Several acoustic measurements on monosyllabic and disyllabic words were performed, and XAB perception tests were performed on sentences. The findings showed that phonetic convergence was not inhibited by dialect mismatch and L1 mismatch between participants and model talkers. The degree of phonetic convergence at the item level (i.e., a given acoustic measurement, such as F0, F2*F1, or vowel duration) was instead influenced by the preexisting phonetic distance regardless of language distance. The larger the preexisting distance, the more native English speakers converged towards the model talkers.

Kim's (2012) study on L1 settings examined phonetic convergence that native English speakers showed towards native English and non-native model talkers. Thus, it is not yet clear if large preexisting phonetic distance facilitates phonetic convergence in L2 settings. The preexisting phonetic distance between L2 learners and native speakers might affect how learners adapt to more native-like pronunciation after being exposed to native speakers. Therefore, it could be the case that the larger the preexisting phonetic distance, the more phonetic convergence could be expected among learners and native speakers as they have more room for acoustic change to happen.

The social and phonetic factors that impact phonetic convergence have been studied mainly in native settings, while their effect in non-native settings remains unclear. For instance, how preexisting phonetic distance between baseline productions and the model talker's productions in a certain acoustic dimension affect phonetic convergence in an L2 setting merits more investigation. Therefore, it would be interesting to examine the correlation between phonetic convergence and the preexisting phonetic distance in such a setting.

1.4.3 Measuring Phonetic Convergence

Previous studies have used several approaches to measure phonetic convergence in speech. Goldinger (1998) was the first to adapt the common AXB perceptual similarity test to evaluate phonetic convergence. This type of perceptual testing has since been widely used to holistically assess phonetic convergence (e.g., Babel & Bulatov, 2012; Kim, 2012; Kim et al., 2011; Namy et al., 2002; Pardo, 2006; Pardo et al., 2017; Pardo et al., 2018; Shockley et al., 2004). In this perceptual task, A and B consist of a talker's production of the same word, phrase, or sentence. A is the production from the pre-test task, while B is the production from the shadowing task. X is the model talker's production of the same word, phrase, or sentence. These productions are presented to native listeners, who are asked to decide whether A or B sounds more similar to X. If listeners report that B is more similar to the model talker's utterance, this indicates phonetic convergence. Listeners base their judgments on holistic perceptions of various factors and features in speech. Therefore, this perceptual task can give reliable results in terms of convergence, divergence, and maintenance; however, this task does not show which phonetic features are affected by phonetic convergence. Thus, the perceptual task does not specify the phonetic features that listeners use to perceive phonetic convergence (Kim, 2012).

Previous studies have conducted acoustic measurements of participants and model talkers' utterances to estimate the degree of phonetic convergence. However, many studies, such as Babel and Bulatov (2011), Pardo (2013), and Pardo et al. (2017), found no single acoustic feature that could predict perceptual phonetic convergence consistently and found that the same talker might show convergence in one feature and divergence in another. Therefore, to obtain a fuller picture of phonetic convergence, it would be necessary to use various acoustic measurements and perceptual similarity judgments in which listeners can use all the available features in the speech signal to detect convergence. One of the acoustic features used in previous studies (e.g., Babel & Bulatov, 2012; Pardo, 2010; Pardo et al., 2012; Pardo et al., 2010) is the F0 of vowels, and another is the voice onset time (VOT) of word-initial stops (Nielsen, 2011). Other studies, such as Babel (2009, 2012), Pardo (2010), Pardo et al. (2012), and Pardo et al. (2017), have measured vowel formants (F1 and F2) to examine phonetic convergence. Vowel duration has also been measured (e.g., Pardo, 2013; Pardo et al., 2012).

1.4.4 Phonetic Convergence in an L2 Setting

Although a large number of studies have investigated phonetic convergence under different social and linguistic conditions among speakers of the same L1, few have examined this issue in L2 settings to find evidence of phonetic convergence to non-native speech (e.g., Gessinger et al., 2020; Ghanem, 2017; Hosseini-Kivanani et al., 2019; Kim, 2009; Kim et al., 2011; Kwon, 2021; Liu, 2017; Olmstead et al., 2021; Rojczyk, 2012; Tobin, 2013; Ulbrich, 2021; Zając & Rojczyk, 2014).

Rojczyk (2012) examined the convergence of Polish learners of English to the vowel /æ/ when shadowing a native English model talker. This vowel was examined because Polish does not have low front vowels. As a result, Polish learners of English have difficulty forming a new category for this English vowel and usually assimilate it to the Polish front mid /e/ and low central /a/. To assess the degree of convergence to the native model talker, the Euclidean distance of learners’ vowels was calculated and compared to that of the model vowels. The findings showed that Polish learners converged their vowel to the model talker. This indicated that when L2 learners were exposed to a native model talker in a shadowing task, they modified their pronunciation of non-native vowels. However, it is not clear how long this modification would last after the shadowing task and how such an effect would help L2 learners improve their pronunciation.

Another study by Zając and Rojczyk (2014) investigated Polish speakers' imitation of the duration of /æ, ɛ, ɪ, iː/ in monosyllabic English words. They showed more convergence in vowel duration towards the native model talker than to the non-native model talker. The authors attributed this pattern to participants' attitudes towards native speakers, as the participants might want to sound native-like. However, participants were neither asked about their attitudes towards native speakers nor about their desire to sound more native-like.

Liu (2017) examined this phenomenon in Mandarin speakers learning English after being exposed to the same model talker, but under tow conditions: a professor and an employer, who shared learners' L1. To evaluate phonetic convergence, the study examined vowel formants and duration. It also examined the duration of word-final stops, and word-initial and word-final consonant clusters. The effect of model talker profession was not strongly supported as participants converged to both in all acoustic measurements in the shadowing task. The author argued that this finding was supported by Automatic Alignment Theory. However, participants showed more convergence to the professor only in terms of the duration of vowels and word-final consonant clusters in the post-shadowing task. That was attributed to a social factor that L2 learners are likely to converge to model talkers when they want to be socially closer to them (in the professor's case). In this condition, the participants were informed that they should make the model talker like them, which made them converge more. In the employer condition, participants were informed that the model talker worked for a company to which they were applying for a job. The study did not explain why the social condition did not affect phonetic convergence in the shadowing task and in the other acoustic features.

There has been growing research on the patterns of phonetic convergence in L2 speech. For instance, Olmstead et al. (2021) examined how Mandarin L2 learners of English changed the production of some difficult English vowels when interacting with English native speakers or Mandarin L2 learners. The aim was to find what patterns of convergence learners showed in these interactions. They examined vowels /i, ɪ/ since native Mandarin speakers have difficulty with the production and perception of this English vowel contrast. The findings indicated that participants converged their vowels differently based on the interlocutor. When they worked with Mandarin speakers, they converged vowel duration, whereas they converged their vowel spectra when interacting with English speakers. However, I am not aware of any research that has examined phonetic convergence exhibited by Arabic L2 learners of English in non-native settings.

1.5 Pronunciation Challenges for Arabic Speakers of English

This study was conducted on Saudi Arabic speakers learning English as it is likely these learners have similar segmental difficulties with English. At the segmental level, Arabic speakers who learn English often err in producing the voiceless bilabial stop /p/ and voiced labiodental fricative /v/ because these do not exist in their L1 phonemic inventory. Instead, Arabic possesses their counterparts: the voiced bilabial stop /b/ and the voiceless labiodental fricative /f/.

Although Arabic and English have a two-way oral stop contrast, their acoustic dimensions differ. For example, Flege and Port (1981) found that word-initial /t/ and /k/ in Saudi Arabic were aspirated but with shorter VOT than the long-lag stops found in English. That indicates that Arabic voiceless stops resemble English voiced stops in terms of VOT (Evans & Alshangiti, 2018). Arabic voiced stops are typically characterized by pre-voicing (or lead VOT). Flege and Port (1981) also found the closure intervals of voiced stops were shorter than those of voiceless stops in word-initial position in Saudi Arabic. English does not have such a temporal contrast. Regarding fricative acoustic features, previous studies (e.g., Crystal & House, 1988a, 1988b; Jongman et al., 2000) have shown that the frication noise for English voiceless fricatives is much longer than their voiced counterparts at a given place of articulation. This temporal contrast also exists in Arabic (Al-Khairy, 2005).

Modern Standard Arabic has a small vowel inventory that consists of three monophthongs with their long counterparts: /u, u:/, /a, a:/, and /i, i:/. Vowel length in Arabic is thus phonologically distinctive. Saadah (2011) investigated the native production of these vowels by Palestinian Arabic speakers. High long vowels had a higher F1 than high short vowels, but short low /a/ had a higher F1 than long /a:/. Interestingly, the high short vowels /i/ and /u/ had equal F1 values while their long counterparts had almost the same values. For F2, /i:/ had higher values than /i/, whereas /u:/ had lower values than /u/. This makes the vowel space for the long vowels bigger than that of the short vowels.

Similar to English, vowel height affects vowel duration in Arabic. That is, low vowels are longer than high vowels (Mitleb, 1984). This can be ascribed to the degree of jaw lowering required in the articulation of low vowels. However, the difference in duration between Arabic long and short vowels is greater than that between English lax and tense vowels.

Given the smaller size of the Arabic vowel inventory, it is expected that Arabic learners of English, which has a far larger vowel inventory, will face difficulty acquiring English vowels that do not have direct Arabic counterparts (Evans & Alshangiti, 2018). For example, most

Arabic learners of English confuse /ɪ/ with /ɛ/, with words like *pen* pronounced like *pin*. Evans and Alshangiti (2018) showed that Saudi Arabic speakers learning English, even with high proficiency, confused several vowels. Based on the vowel identification task reported by British English speakers, the vowels that "were highly confusable are /ɜː/-/eə/, /ɒ/-/ʌ/, /ʊ/-/əʊ/-/uː/, /eɪ/-/aɪ/ and /ɪ/-/e/" (p. 27).[1] These learners will likely have the same difficulty with similar American English vowels. However, only certain sounds were chosen for the training sessions to make the task more manageable and to avoid overwhelming the participants. The consonants /p/ and /v/, the monophthong /ɛ/, and the diphthong /oʊ/ were used in nonsense words in the training sessions as they were predicted to be difficult to Arabic L2 learners of English.

1.6 The Aim of the Study

This study sought to extend the current literature on what voices L2 learners should imitate to improve their pronunciation. It explored how phonetic convergence towards native speakers could be used to improve the pronunciation of Arabic learners of English. More specifically, it explored the extent to which phonetic convergence might help learners improve their pronunciation of English segments that were expected to be difficult.

The study also investigated phonetic factors that might affect phonetic convergence towards native speakers in L2 settings. In particular, it scrutinized how the preexisting phonetic distance between learners and native model talkers affected phonetic convergence. It was expected that the larger this distance, the more phonetic convergence there would be among the learners and native speakers.

[1] The IPA transcription of these vowels reflect the British English pronunciation: /ɜː/ *heard* and /eə/ *haired*, /ɒ/ *hod* and /ʌ/ *hud*, /ʊ/ *hood* and /əʊ/ *hoed* and /uː/ *who'd*, /eɪ/ *hayed* and /aɪ/ *hide*, /ɪ/ *hid* and /e/ *head*.

Another aim was to investigate the benefits of shadowing as a technique for learning L2 pronunciation. The study examined whether shadowing was an effective way to improve segmental intelligibility in general or was constrained by learners' ability to phonetically converge to their trainers. It was therefore a goal to see whether shadowing would be affected by learners' ability to converge more to some native speakers or would be effective even when L2 learners did not exhibit phonetic convergence.

The study sought to answer the following research questions:

1. Do Arabic learners of English improve their segmental productions more when they are trained by a native model talker to whom they phonetically converge?
2. Does the preexisting phonetic distance between native model talkers and Arabic learners of English determine the degree of convergence?
3. Is the shadowing technique generally effective in improving pronunciation, or is it constrained by L2 learners' degree of phonetic convergence?

1.7 Significance of the Study

This study is important to the field of speech perception because it provides insight into how speech perception affects speech production in an L2 setting. Finding evidence for phonetic convergence towards native speakers in this setting would support the current models of speech perception and production. Motor Theory by Liberman and Mattingly (1985) and exemplar-based models like those proposed by Johnson (1997, 2006) and Pierrehumbert (2001, 2003) suggest a strong link between speech perception and production. Thus, phonetic convergence to native speakers would indicate that L2 learners are capable of hearing L2 phonetic details and modifying their speech accordingly after being exposed to native speech.

The study will also help find which phonetic factors affect phonetic convergence in an L2 setting and to what extent L2 learners use phonetic convergence to improve their pronunciation. In their review paper, Nguyen and Delvaux (2015) discussed the role phonetic convergence plays in the development of phonological systems. They argued that when interlocutors interact and converge phonetically to one another, their production is affected, and as time passes, phonetic convergence accumulates. They added that phonetic convergence could be a driving mechanism that L2 learners use to acquire L2 phonetics and phonology. However, they based their claims on reviewing papers that examined gestural drift (e.g., Sancier & Fowler, 1997) that did not address how L2 learners utilize phonetic convergence to develop L2 phonetic and phonological systems. This study examined the relationship between phonetic convergence and L2 pronunciation learning more rigorously.

It is important to see whether shadowing as a learning tool contributes to pronunciation improvement and is effective in all training sessions or if it is constrained by other factors, such as learners' ability to converge more to some native speakers but not to others. The study examined what native speaker voices L2 learners should imitate to improve their production. The study examined whether it is likely that shadowing with "suitable" native speakers based on phonetic convergence helps L2 learners create segmental and suprasegmental representations in long-term memory, not only during shadowing.

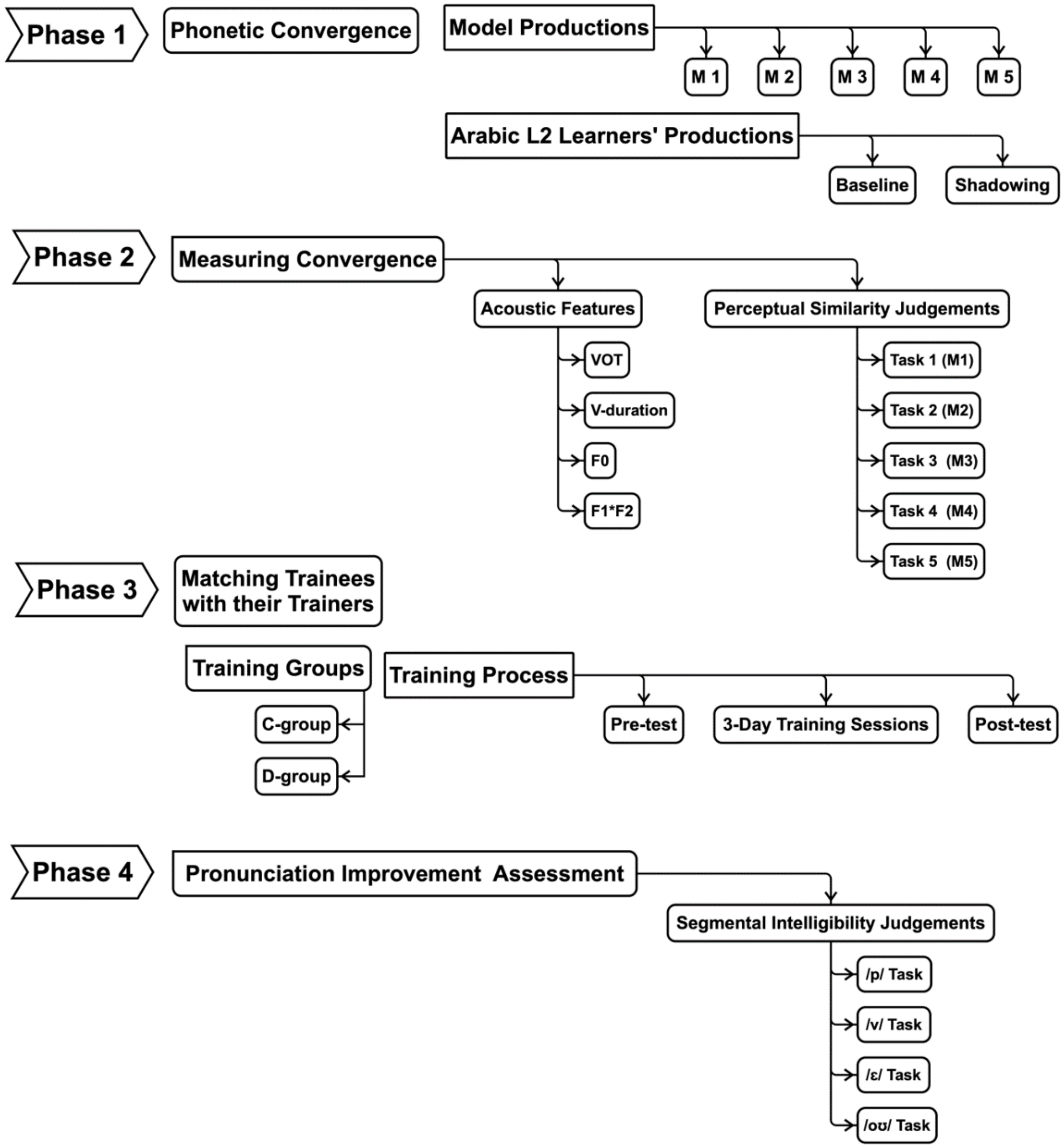
CHAPTER 2: GENERAL METHODOLOGY

2.1 Methodology

This chapter explains the methodology of the study. To accomplish the goals laid out in the previous chapter, the study went through several phases, as illustrated in Figure 1. The first phase involved measuring phonetic convergence among native English speakers and Arabic speakers learning English. To measure phonetic convergence, learners were asked to provide baseline productions of a list of English words chosen according to criteria discussed in Chapter 3. Then, each learner shadowed five model talkers saying the same list of words the learners had read to obtain their baseline productions. Learners shadowed the five model talkers in five different blocks on the same day.

Some studies have employed more than one model talker shadowed by the same participants on the same day, but the stimuli were blocked by model voice. For example, in Namy et al. (2002), male and female participants shadowed four model talkers (two men and two women) in four different blocks. They found that female participants converged more than male participants, and they converged more to male model talkers. Another study by Babel and McGuire (2014) had five model talkers that participants shadowed under two conditions: high variability and low variability. Participants in the high-variability condition were presented with the words read by the five model talkers in random order, whereas the ones in the low-variability condition shadowed each model talker in a separate block. More convergence was found in the low-variability condition. None of these studies reported that phonetic convergence was negatively affected when participants shadowed different model talkers in different blocks.

Figure 1. *Schematic Diagram of the Experimental Design of the Study*



After all the recordings were segmented manually at word and phoneme levels, two methods were used to evaluate phonetic convergence: perceptual tasks and acoustic measurements. Five online perceptual similarity judgment tasks were created using PsyToolkit (Stoet, 2010, 2017) that were assessed by native English speakers recruited from Prolific.co, an online platform that hires people to take part in studies. This online platform was used to collect data for perceptual and intelligibility judgments to avoid social contact with judges during the pandemic.

With regard to acoustic attributes, values of VOT, vowel duration, F0, and F1*F2 were measured. These values were converted to difference-in-distance (DID) scores, which estimated the effect of model talker on shadowed productions by comparing baseline differences between each shadower and model talker as well as shadowed differences. Positive DID values showed convergence in shadowing tasks, and negative values indicated phonetic divergence. Values of zero displayed no change (maintenance) of the baseline values. The details of perceptual similarity judgments and acoustic measurements are explained in Chapter 3.

Based on the degree of convergence among learners and native model talkers, the learners were assigned to two training groups, as explained in Chapter 4. The assigning criteria were based mainly on the similarity perceptual judgements, that is, how L2 learners were perceived as more similar to or different from the model talkers. Acoustic convergence was considered only when a learner obtained the same proportion for two or more native English model talkers. Thus, the first group, referred to as C-group (convergence group), was trained by the model talkers to whom they showed the highest degree of phonetic convergence, whereas the second group, D-group (divergence group), received training from model talkers who they diverged from or showed the least convergence to. The same native model talker could be a trainer in both groups based on the convergence or divergence that each learner showed. It is worthwhile to note that each learner was trained by one native model talker following low-variability training to avoid overwhelming them with more than one trainer. The motive for using a low-variability paradigm is explained below.

Many studies (e.g., Barriuso & Hayes-Harb, 2018; Bradlow et al., 1997; Iverson & Evans, 2009; Lively et al., 1993; Nishi & Kewley-Port, 2007; Perrachione et al., 2011; Pisoni & Lively, 1995) have shown that high-variability phonetic training (HVPT) has more efficacy in learning non-native sounds compared to low-variability phonetic training (LVPT). HVPT refers to a training approach that exposes learners to acoustically varied speech produced by multiple talkers in various phonetic contexts. However, in LVPT, learners are trained by a single talker. Researchers have shown that HVPT is effective in improving the ability to perceive non-native sounds. Bradlow (2008) argued that when learners are exposed to HVPT stimuli, their acquisition of non-native contrasts is promoted. Exposing learners to multiple speakers appears to have great effectiveness in generalizing perceptual learning in novel speakers, and HVPT contributes to the formation of robust perceptual categories. That means learners can apply these categories to novel talkers and novel phonetic environments. HVPT thus enables learners to make generalizations of what they have learned to untrained words and new talkers.

Long-term retention of new phonetic contrasts is another strength of HVPT, which has been found to help learners maintain what was learned after an extended period of time without more training or feedback (Pisoni & Lively, 1995). Therefore, HVPT leads to long-term retention of new perceptual categories without additional training.

Although HVPT offers more benefits than LVPT, the production training sessions in this study followed low-variability procedures. That is, each participant was trained by one native model talker. The main interest in this study was whether L2 learners would show greater improvement in their L2 pronunciation when trained by native model talkers to whom they exhibited more convergence. In other words, the focus was on the relationship between phonetic convergence and the acquisition of L2 segments.

Phonetic convergence is negatively affected by talker variability in native settings. Bable and McGuire (2015) found that when speakers shadowed several model talkers, the degree of convergence decreased during the experiment. However, when speakers shadowed one model talker, they showed more constant convergence throughout the shadowing task, and the imitation increased slightly as they were exposed more to the model talker. Bable and McGuire attributed the decline in phonetic convergence in high-variability contexts to increased cognitive load, which leads to "less detailed sensory analyses" (p. 4). Therefore, using the HVPT paradigm in this study would make it harder to draw direct conclusions about the role of phonetic convergence in the acquisition of L2 segmental structures. Furthermore, investigating generalization and long-term retention was beyond the scope of this study. Thus, production training sessions with high-variability components should lessen L2 learners' ability to converge to model talkers, and less improvement would be expected in their pronunciation. Taken together, it was important to use a low-variability paradigm in this study.

After matching the trainees with their trainers, there were three training sessions in which trainees were asked to learn nonsense words. The trainees in both groups followed the same procedure. They were presented with 12 English nonsense words divided into two sets. One set included the target consonants /p, v/, and the other contained the target vowels /ɛ, oʊ/. More details about the criteria to build these words are presented in Chapter 4. The same nonsense words were used in the pre-test, training, and post-test.

For the pre-test, trainees were asked to produce the nonsense words presented visually along with an image of non-objects on a screen. In the training sessions, trainees were presented with a picture, then an audio stimulus indicating the picture's meaning. Their task was to repeat the word they heard as quickly and clearly as possible. No further instructions or feedback were provided. The aim was to see whether phonetic convergence would facilitate learning the target word's pronunciation and to what extent learners could improve their segmental pronunciation without explicit instructions on how these words were pronounced.

The final stage of the study was the assessment of pronunciation improvement on the post-test. Four native listeners' judgment tasks, two for the consonants and two for the vowels, were constructed using PsyToolkit (Stoet, 2010, 2017). Each of these experiments was presented to a different group of 10 native English speakers to evaluate the intelligibility of the examined segments. Forty native English listeners (10 for each task), recruited from the Prolific platform, assessed the intelligibility of the training segments from the pre- and post-tests. Native listeners' judgments are important to understand how intelligibility can affect communication and to what extent L2 learners can be understandable (Edwards & Zampini, 2008). Munro and Derwing (1995) defined intelligibility as "the extent to which an utterance is actually understood" (p. 291). Intelligibility measures how understandable a speaker's speech is at the word or sentence level. Thus, this study examined whether a given degree of phonetic convergence would lead to pronunciation improvement in intelligibility after phonetic production training. The intelligibility judgements are explained in Chapter 4.

2.2 Statistical Analysis

All statistical tests were conducted using SPSS (Version 27.0). Numerous parametric and non-parametric tests were carried out to determine the relationships between variables. In XAB perceptual tasks, the reliability of the English raters was tested by computing two reliability tests: Fleiss' Kappa (Cohen, 1960) and intraclass correlations (Shrout & Fleiss, 1979). Then, based on the results of a Kolmogorov-Smirnov test of normality, one-sample *t*-tests and one-sample

Wilcoxon signed rank tests were conducted on averaged proportions of selected shadowed items to examine whether listeners' performance was above chance level. A Kruskal-Wallis test was also used to find any significant differences between the five XAB tasks. Post-hoc comparisons were carried out using the Mann–Whitney test with the Bonferroni correction.

In acoustic convergence, one-sample *t*-tests and one-sample Wilcoxon signed rank tests were used to examine if the DID values were significantly higher than zero. Within the same acoustic feature, a one-way repeated measures ANOVA and Friedman's two-way analysis of variance by ranks were conducted to compare DID values across the five model talkers.

To determine a relationship between preexisting phonetic distance and degree of convergence, generalized estimating equation (GEE) models with an identity link function were carried out. These models were introduced by Liang and Zeger (1986) as an extension of generalized linear models. They can handle correlated longitudinal, repeated, and cluster data and enable regression analysis on data that do not have a normal distribution (McCullagh & Nelder, 1989). GEE modeling was also used to capture the influence of acoustic convergence on raters' evaluation of phonetic convergence.

In the training experiment, differences between the two groups in terms of acoustic and perceptual convergence were examined using independent-samples *t*-tests. One-sample *t*-tests were used to examine whether the DID values of each acoustic feature in the two groups were significantly higher than zero. Fleiss' Kappa and intraclass correlations were used to assess the reliability of the English raters in the segmental intelligibility judgments. To compare both trainee groups' performance on the pre- and post-tests, a mixed between/within-subjects analysis of variance (a mixed-design ANOVA) was used. If any violation of the mixed-design ANOVA was detected, the Mann–Whitney test was used as the alternative non-parametric test. The last statistical analysis was GEE modeling to capture the impact of phonetic convergence in acoustic features on the magnitude of change from pre- to post-test.

CHAPTER 3: PHONETIC CONVERGENCE

3.1 Phonetic Convergence

This chapter reports the patterns of phonetic convergence exhibited by learners in the experiments. First, it discusses the design of the phonetic convergence experiments, including the native English model talkers, Arabic participants, stimuli, English listeners, and data analysis. It details how phonetic convergence was measured acoustically and perceptually as judged by native English listeners. It then demonstrates the relationship between the preexisting phonetic distance between L2 learners and native model talkers and the degree of convergence in acoustic features. The last section reports the relationship between acoustic convergence and the perceptual similarity judgments.

3.2 Participants

The participants were 26 female Saudi native Arabic speakers whose ages ranged from 20 to 47 (*M* = 28.42, *SD* = 6.12). All of them had studied English as a foreign language since the seventh grade in intermediate school (from about age 13) in Saudi Arabia. Their native Saudi dialects were mainly Hijazi (10 participants) and Najdi Arabic (11). The other five were speakers of Eastern Saudi Arabic (three) and Southern Saudi Arabic (two). Although there are some dialectal differences among these varieties, the difficulties they face in learning English as an L2 are similar. Additionally, L1 transfer patterns appearing before the training sessions were expected to be the same for each speaker. A detailed description of L1 transfer is given in the training stimuli. These participants were recruited from the University of Wisconsin-Milwaukee, Cardinal Stritch University, and Marquette University. All of them volunteered to take part in the research.

No participant reported any history of speech or hearing impairment. Their length of residency in the U.S. ranged from two to 10 years (*M* = 5.27, *SD* = 2.46), and all of them had studied English as an L2 in the U.S. for at least two semesters. Their International English Language Testing System (IELTS) scores ranged from 5 to 6.5 (*M* = 6), which would be considered intermediate to high-intermediate level, respectively. This study was conducted on learners with a high-intermediate level to ensure they would not struggle with basic English sounds.

3.3 Model Talkers

Five female native English speakers served as the model talkers to be shadowed. They all spoke Upper Midwestern American English (Purnell et al., 2017), were recruited from the University of Wisconsin-Milwaukee, and ranged in age from 19 to 23 (*M* = 20). No model talkers reported any speech, language, or hearing disorder. They were compensated with $10 for their participation.

3.4 Stimuli

Twelve low-frequency disyllabic words were used to assess phonetic convergence. According to Pardo et al. (2017), phonetic convergence is more evident in disyllabic words because they are longer in duration and allow more opportunity for convergence. Following Goldinger (1998), low-frequency words were chosen to evoke more convergence. Following Kim (2012), the criterion for determining word frequency was a maximum of 30 per million words in a corpus. In this study, two corpora were consulted to find word frequency: SUBTLEXus (Brysbaert & New, 2009) and the Corpus of Contemporary American English (COCA) (Davies, 2008). The list of words with their exact frequency is shown in Table 2. As shown in the table, the frequency of all words in both corpora was under 10 per million, even less than the frequency Kim (2012) used in her study.

Table 2. *Word List Used for Baseline and Shadowed Productions*

V1	C1	C2	Frequency per million	
		[voiced]	COCA	SubtlexUS
/i/	t	teaser	0.87	0.27
	d	diesel	7.51	2.65
	k	keeler	0.63	1.67
	g	geezer	0.45	1.35
/æ/	t	tagger	0.05	0.12
	d	dagger	2.76	4.92
	k	cabbage	6.39	2.90
	g	gadget	2.20	2.43
/ʌ/	t	tugging	2.73	0.41
	d	dubbing	0.01	0.02
	k	cudgel	0.27	0.02
	g	guzzler	0.10	1.39
Mean			1.9975	1.5125

All the words were disyllabic with stress on the first syllable, and the syllable structure was either CVCVC or CVCCVC to avoid stress transfer from Arabic, as the first syllable in Arabic is stressed in such structures. Thus, learners were expected to assign stress to the first syllable because they acquired the correct pattern of stress in these English words or they transferred their L1 stress rules. That helped control the stressed vowels, so native speakers and L2 learners were able to produce the vowel in the same manner.

The first syllables in all words had the target vowels. There were four instances of [i, æ, ʌ], which were chosen for two reasons. First, based on Evans and Alshangiti (2018), Saudi learners of English with a proficiency higher than novice have no difficulty producing these vowels. Second, based on Althubyani and Park's (2019) research on phonetic convergence in L2 settings, Saudi speakers learning English show the most convergence in these vowels.

Since VOT was measured, I had the only word-initial consonants be the alveolar and velar stops /t, d, k, g/. There were three words for each target stop that occurred word-initially followed by the target vowels that were analyzed. The second consonant in all words was controlled for voicing (i.e., the second consonant was voiced). The words used in this study were selected by taking into consideration the pronunciation difficulties that learners might face. For this reason, /p/ was excluded from the list since it is problematic for Arabic speakers. Moreover, vowels such as /ɛ/ and /ʊ/ were excluded since Arabic speakers, particularly Saudis, tend to confuse these vowels with /ɪ/ and /u/, respectively (Evans & Alshangiti, 2018). Words that had nasal stops after the first vowel were also avoided since Arabic speakers, particularly the learners in this study, do not nasalize vowels before nasals. Thus, there might be some difference between native English speakers and Arabic speakers when producing English vowels before nasals. For this reason, nasal stops could cause inaccuracies in measuring the difference in distance. In addition, any words with final super-heavy syllables were excluded to avoid stress transfer from Arabic, as final super-heavy syllables always attract stress in Arabic.

3.5 Phonetic Convergence Task Procedures

In the phonetic convergence experiment, participants were asked to do two tasks. The first was a production pre-test in which they read a list of 12 low-frequency disyllabic words that were analyzed to measure their baseline productions. The same word list was presented in five shadowing blocks. Each block had the stimuli spoken by the same native model talker. However, to avoid familiarizing learners with the target words used in each shadowing block, five different sets of filler words were used in each block. That made the words to shadow less predictable in the following block and made learners pay more attention to what they heard in every block. There was a break after each block. For each learner, the order of the blocks was randomized. The participants were recorded individually in a quiet room using Audacity (Version 2.4.2) at the University of Wisconsin-Milwaukee. Soundproof booths were not used due to circumstances related to the pandemic. All spoken materials were recorded on a personal laptop connected to Focusrite Scarlett Solo Audio Interface (3rd Gen) through a UHURU mounted cardioid microphone with a sampling rate of 44,100 Hz. To minimize unwanted background noise, a microphone isolation shield was used.

The second task was shadowing native speakers as they said the same list of words. During the shadowing exercise, participants were asked to repeat the words they heard as quickly and clearly as possible after a model talker. Neither explicit instructions to imitate the model nor the script of the list of words were provided.

3.6 Perceptual Assessment: XAB Perceptual Similarity Judgements

3.6.1 Listeners

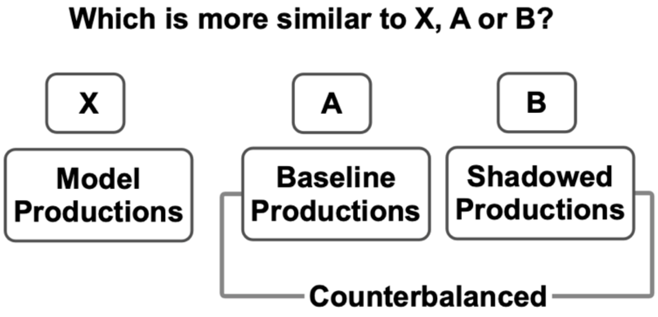
The researcher recruited 55 native English listeners via Prolific to participate in XAB perceptual similarity tasks (11 per task). Out of the 55, 34 (78%) were female, and their ages ranged from 19 to 65 (*M* = 37.93, *SD* = 11.52). They were paid $8 for their participation. Listeners were prescreened to be English-speaking monolinguals born in the U.S. Listeners who missed attention checks were excluded from the study. As the data were collected online via Prolific, attention checks were crucial to detect careless responses and ensure listeners were paying attention throughout the task. None reported any hearing or speaking impairments in the survey provided before they started the task.

3.6.2 Procedures

In AXB perceptual-similarity tests, if the shadowed productions of a given speaker are perceived as more similar to the model productions, it means the speaker exhibits phonetic convergence. Following Kim et al. (2011), an XAB (rather than an AXB) discrimination task was used since it creates less memory load on participants in the perceptual judgment task. Using PsyToolkit (Stoet, 2010, 2017), five perceptual similarity judgment tasks were designed. English listeners were presented with three repetitions of the same word (XAB). X here represents the model productions, whereas A and B are the baseline and shadowed productions taken from the same participant. Listeners were asked to determine if A or B sounded more similar to X. That is, they were to decide which of the learner's productions sounded more like the native model talker. Listeners responded by pressing one of two keys corresponding to A or B that appeared on the computer screen. The first box (X) was inactive. The presentation of A and B were counterbalanced with the native model talker's productions at the beginning. Figure 2 illustrates how XAB perception tests were designed and presented to the listeners. If learners converged to the native model talkers, their shadowed productions should have sounded more like the native model talker productions (X) than the baseline productions that were recorded before their exposure to the model talkers.

Five XAB tests were created to make the task length manageable for the listeners and to reduce the number of trials presented to them. In each XAB test, 11 listeners assessed the similarity between one model talker and 26 Arabic participants. Thus, there were 11 judgments for each model/learner pair. Each XAB had 26 blocks for the 26 learners to keep the shadower productions A and B the same throughout the block. In each block, there were 24 trials (12 words x 2 repetitions) in which the presentation of A and B was counterbalanced. In each trial, there were three presentations of the same word (XAB), where X was the model talker production, and A and B were the baseline and shadowed productions. Thus, each XAB had 624 trials presented in 26 blocks (26 learners x 12 words for one model talker x 2 repetitions).

Figure 2. *Depiction of XAB Perceptual Similarity Tests*



Each XAB compared only one model talker’s productions with all learner productions before and after exposure to that model talker. In total, 1,932 words were extracted to be presented to the listeners in the five tasks: (5 model talkers × 12 words) + (26 learners × 12 words × 6 (5 shadowed + 1 baseline) productions). Since each word was repeated twice, each XAB task had 1,872 words in total: 24 trials (12 words x 2 repetitions) x 3 (model + baseline + shadowed) x 26 blocks. Thus, the total number of words presented to listeners were 9,360 (1,872 x 5 XAB tasks). The XAB tests are illustrated in Table 3. Listeners had the choice to take a break after five blocks. To make sure listeners were paying attention to the task, and not only pressing A or B randomly, four attention checks repeated twice were added to each task. The category options (A or B) were the same within each task, but there was no audio file played. A simple question appeared on the screen asking about the first letter of the words. They were given words starting with A or B, such as "apple" and "bank," and their task was to select A or B. The inter-sample interval was 600 milliseconds, and the inter-trial interval was 1,000 milliseconds before the next trial started. Each word from each learner was presented twice to the perceptual judges. The entire task took around 45–60 minutes to complete.

Table 3. *Summary of XAB Tests*

XAB tests	One for each native model talker Task 1: Model Talker 1 Task 2: Model Talker 2 Task 3: Model Talker 3 Task 4: Model Talker 4 Task 5: Model Talker 5
Blocks within a single XAB test	26 (one for each learner)
Trials within each block	24 (12 words x 2 repetitions)
Stimuli in each trial	3 (XAB): X = native model production, A and B = baseline and shadowed words counterbalanced

3.6.3 Results

As discussed in the previous section, each XAB compared only one model talker's productions with all learner productions before and after exposure to that model talker. For each model talker, 11 unique judges were assigned to rate participants' productions, resulting in 624 ratings per judge for a total of 11 (judges) x 624 (trials) = 6,864 individual ratings. Each judge was presented with the same number of trials and category options for each trial (A or B). First, inter-rater reliability was calculated for each of the five subsets using Fleiss' Kappa to see if the probability of raters' agreement was significantly above chance levels. Fleiss' Kappa is an appropriate measure of agreement at this level because the judge ratings are nominal and binary (Cohen, 1960). In this way, the probability of agreement was shown to be above chance level (i.e., 0.50), and overall agreement between judges for all five subsets was above chance level ($p < .001$).

To test the reliability of judges' ratings and to ensure there were no significant inter-rater differences, intraclass correlations (Shrout & Fleiss, 1979) were computed using SPSS for each of the five subset tasks (XAB). This type of reliability statistic shows if the raters are consistent with one another. First, the researcher calculated the number of trials on which a shadowed production was selected as more similar to the model talker's production. Then, the proportion of times judges chose shadowed words compared to the total number of selections was calculated independently for each judge across each subject and within each task. For a single task, the new variables resulted in 26 values (one per subject) for all 11 judges within the task. To measure inter-rater reliability between judges for the proportions of shadowed words, the intraclass correlation was calculated individually for each task. Tasks 1 and 5 showed good consistency among judges. The correlation of coefficients for Task 1 averaged .82 with a 95% confidence interval from 0.69 to 0.91 ($F(25, 250) = 5.449$, $p < .001$), while Task 5 averaged .81 with a 95% confidence interval from .69 to .90 ($F(25, 250) = 5.37$, $p < .001$).

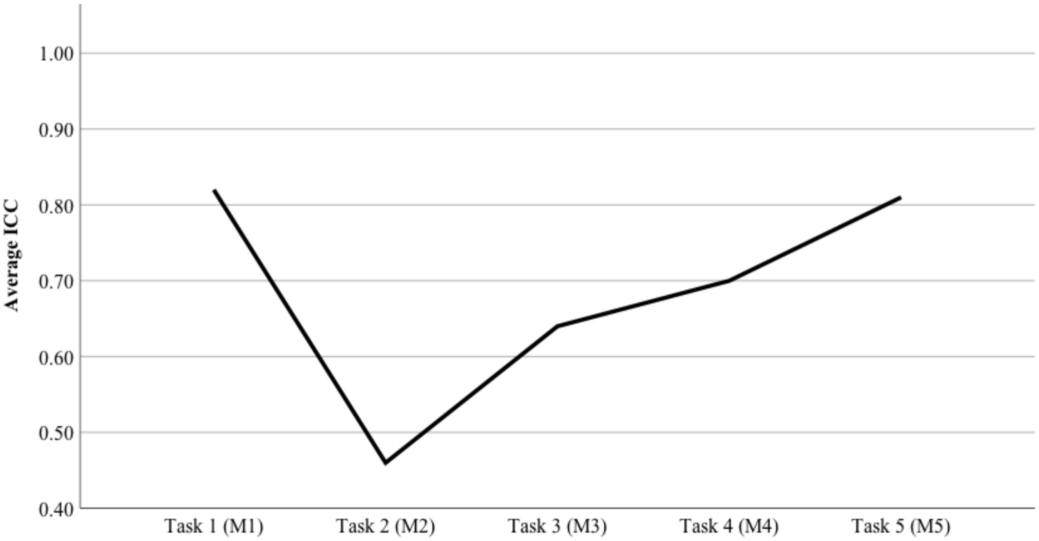
Task 2 had poor internal consistency (ICC = 0.46), which was the lowest among the tasks. Table 4 shows the reliability statistics for each task, and Figure 3 demonstrates the intraclass correlation coefficient value for each task. In the figure, M represents the model talkers. Judges assessed convergence between one model talker and learners in each task.

Table 4. *Intraclass Correlation Coefficient for Each XAB Task*

	ICC	95% Confidence Interval					
		Lower Bound	Upper Bound	Value	*df1*	*df2*	*p*
Task 1 (M1)	0.82	0.69	0.91	5.45	25	250	< 0.001
Task 2 (M2)	0.46	0.10	0.72	1.86	25	250	< 0.001
Task 3 (M3)	0.64	0.39	0.82	2.89	25	250	< 0.001
Task 4 (M4)	0.70	0.49	0.84	3.29	25	250	< 0.001
Task 5 (M5)	0.81	0.69	0.90	5.37	25	250	< 0.001

Note. ICC = average intraclass correlation coefficient, 95% Lower/Upper CI = 95% confidence interval of ICC, and M is an abbreviation for model talker.

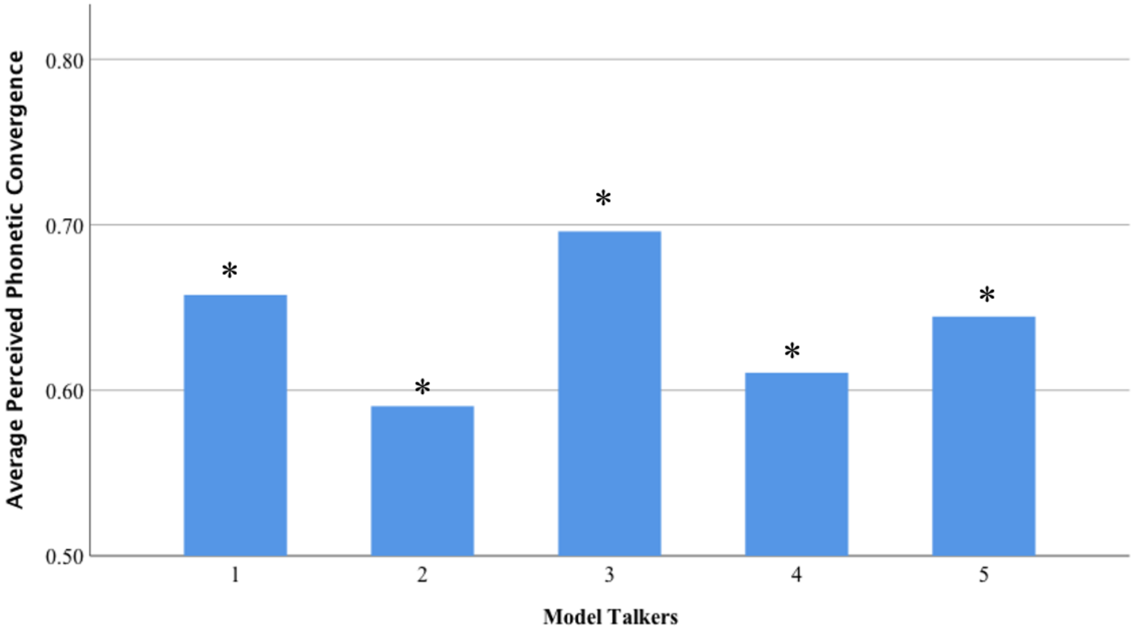
Figure 3. *Average Intraclass Correlation Coefficient by XAB Task*



Next, one-sample *t*-tests were conducted on averaged proportions of selected shadowed items to determine if listeners' performance was above chance level (0.5). Tasks 1–4 were normally distributed; therefore, one-sample *t*-tests were performed. The averaged proportions of selected shadowed items in the four tasks were significantly higher than 0.5: Task 1 (*M* = .66, *SD*

= .09, *t*(25) = 8.49, *p* = .000), Task 2 (*M* = .59, *SD* = .05, *t*(25) = 9.24, *p* = .000), Task 3 (*M* = .70, *SD* = .05, *t*(25) = 19.00, *p* = .000), Task 4 (*M* = .61, *SD* = .08, *t*(25) = 7.50, *p* = .000). For Task 5, a one-sample Wilcoxon signed rank test, an alternative non-parametric test to one-sample *t*-tests, was used because the data were not normally distributed. The median proportions of selected shadowed items in Task 5 were significantly higher than 0.5 (*Md* = 68, *Z* = 4.25, *p* = .000). These results suggested that native English listeners chose shadowed items as more similar to the model items than the baseline in the five XAB tasks. Figure 4 illustrates the proportions of perceived phonetic convergence across the five model talkers. The asterisk indicates the proportion of shadowed words selected as more similar to the native model talkers was significantly higher than chance level (0.5).

Figure 4. *Average Perceived Phonetic Convergence across Model Talkers*



XAB descriptive statistics by model talker is displayed in Table 5. A Kruskal-Wallis test compared the differences in perceived convergence across the five model talkers. A significant difference was found in perceived phonetic convergence between at least two model talkers (χ^2 (4, 130) = 31.114, *p* = .000). Post-hoc comparisons using the Mann–Whitney test with the Bonferroni correction (.05/10 = .005) revealed a significant difference between Model Talkers 1 (*Md* =.66, *n* = 26) and 2 (*Md* = .60, *n* = 26), *U* = 171.5, z = -3.048, *p* = .002. Model Talker 2 was significantly different from Model Talker 3 (*Md* =.70, *n* = 26), *U* = 47, z =-5.327, *p* =.000, and Model Talker 5 (*Md* =.68, *n* = 26), *U* = 169.5, z = 520.5, *p* = .002. There was also a significant difference between Model Talkers 3 and 4 (*Md* =.63, *n* = 26, *U* = 122.5, z = -3.95, *p* =.000). No other pairwise comparisons were statistically significant.

Table 5. *XAB Descriptive Statistics by Model Talker*

XAB Model Talker	*M*	*SD*	*Md*	*K-S*
1	0.66	0.09	0.66	0.200
2	0.59	0.05	0.59	0.200
3	0.70	0.05	0.70	0.200
4	0.61	0.08	0.63	0.200
5	0.64	0.09	0.68	0.006*

Note. SD = standard deviation of the mean, *Md* = median, and *K-S* = *p*-value for the Kolmogorov-Smirnov test of normality, with *K-S* = *p*-value < .05 meaning the data are not normally distributed.

The XAB perceptual similarities tasks suggested that participants on average converged to all five model talkers. However, some model talkers evoked more convergence than others. This indicated that, in L2 settings, Arabic L2 learners of English can phonetically converge to native speakers of English, and their convergence can be detected by native English listeners.

3.7 Phonetic Convergence on Acoustic Attributes

3.7.1 Analysis

To evaluate phonetic convergence acoustically, a couple of measurements were done using Praat software (Boersma & Weenink, 2021). Prado et al. (2017) indicated that vowel duration and F0 were equally strong in predicting phonetic convergence, followed by vowel spectra. Therefore, F0, vowel duration, and F1 and F2 of first-syllable vowels were measured. In addition, the VOT of the word-initial stops /t, d, k, g/ was measured. First, manual segmentation at phoneme and word level were done on the recordings of the model talkers and learners (baseline and shadowed productions). Next, Praat scripts were run to extract the target acoustic measurements. F0, F1, and F2 values were taken at the midpoint of the vowel duration. Vowel duration was measured from the beginning of the vowel periodicity to the end of clear formant structures. F1 and F2 were obtained at the midpoint of the vowels. Finally, the VOT was measured from the burst release of the word-initial stop until the periodicity of the vowel began. Vowel duration and VOT values were measured in milliseconds.

The values of F0, vowel spectra (F1*F2), vowel duration, and VOT were converted to difference-in-distance (DID) scores, comparing baseline differences between each shadower and model talker as well as shadowed differences. After that, absolute values of the differences for shadowed items were subtracted from absolute values of the differences for baseline items, producing the DID values (DID = baseline distance – shadowed distance). Hence, values greater than zero indicated acoustic convergence due to smaller differences during shadowing than during baseline productions, while negative values indicated divergence, as the distance between the baseline and shadowed productions increased (Prado et al., 2017). Since vowel formants have two dimensions (F1 and F2), the Euclidean distance equation was used to compare each learner's productions to those of the model talkers (Prado at el., 2017). The equation for finding the Euclidean distance is the following:

$$\sqrt{(\text{Participant F1} - \text{Model Talker F1})^2 + (\text{Participant F2} - \text{Model Talker F2})^2}$$

3.7.2 Results

The recorded words spoken by the five model talkers were analyzed in Praat. The total number of words analyzed were 1,932: (5 model talkers × 12 words) + (26 participants × 12 words) × 6 productions (5 shadowed + 1 baseline). Since F0, VOT, vowel duration, F1, and F2 were measured for each word, this yielded 9,660 measurements in total (1,932 words × 5 measurements). The values of each acoustic measurement for the five model talkers were measured (see Table 6, where the number in parentheses is the standard deviation).

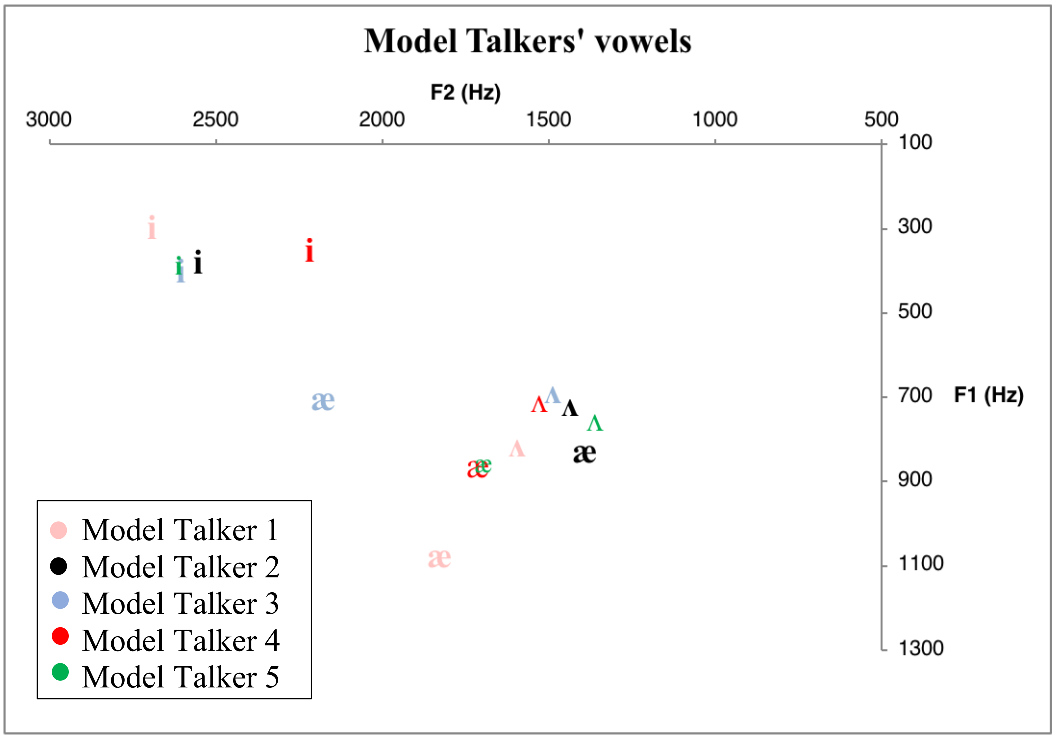
The DID values of F0, F1*F2, vowel duration, and VOT were calculated for each learner with each model talker. First, all DID values were tested to see if they were normally distributed using the Kolmogorov-Smirnov test of normality. If the DID values of an acoustic attribute were normally distributed, one-sample *t*-tests were used to determine whether the DID values were significantly different from zero. However, if the DID values were not normally distributed, the one-sample Wilcoxon signed rank test was used as an alternative non-parametric test.

Table 6. *Mean Values for Each Acoustic Dimension for the Five Model Talkers*

Model Talker	VOT (ms)	V-duration (ms)	F0 (Hz)	F1 (Hz)	F2 (Hz)
1	67.09 (42.97)	152.44 (36.99)	218.71 (12.04)	727.16 (342)	2038.48 (506)
2	61.91 (41.41)	168.84 (43.83)	221.20 (12.25)	639.74 (202)	1793.99 (676)
3	76.71 (49.55)	156.01 (49.20)	179.39 (59.10)	595.46 (186)	2089.87 (616)
4	42.77 (26.75)	154.06 (42.01)	210.75 (15.98)	640.12 (232)	1819.71 (341)
5	58.50 (42.54)	145.01 (29.10)	182.37 (15.50)	666.40 (217)	1889.96 (610)

Before detailing the patterns of phonetic convergence in the acoustic features, I briefly discuss the general findings of how learners converged their English vowels towards the five model talkers. The average values of F1 and F2 for the three English vowels /i, æ, ʌ/ for each model talker are displayed in Figure 5. The means were calculated based on the vowel formant frequency measurements obtained at the midpoint. Although all the model talkers reported speaking an Upper Midwest dialect of English, there are obvious differences between their vowel formant frequencies. Figure 5 clearly shows that Model Talker 4 produced /i/ further back indicated by a lower F2 compared to the other model talkers. The vowel /æ/ produced by Model Talkers 4 and 5 occupy nearly the same positions. However, Model Talker 1 exhibited an /æ/ with a higher F1 than the other model talkers, while Model Talker 3 produced /æ/ with the highest F2. The central vowel /ʌ/ occupies similar relative positions with more consistent vowel formant values across model talkers.

Figure 5. *Model Talkers' Mean Formant Values for Each Vowel from Shadowing Task Stimuli*



Figures 6–10 display the averaged formant frequencies for the three vowels that were used to measure phonetic convergence. They show the direction of phonetic convergence of all participants to each of the five model talkers separately. Note that the average values of vowel formants were used. Model talker vowels are plotted in dark orange, learner baseline vowels are plotted in pink, and shadowed vowels are plotted in blue. Learners converged the majority of their vowels to all model talkers, even though the degree of convergence varied. For example, the low vowel /æ/ exhibited the highest degree of convergence when learners shadowed Model Talker 3 and lower convergence to Model Talker 1. However, the same vowel /æ/ showed divergence from Model Talkers 2 and 5.

Figure 6. *Formant Plot Displaying Phonetic Convergence of All Learners to Model Talker 1*

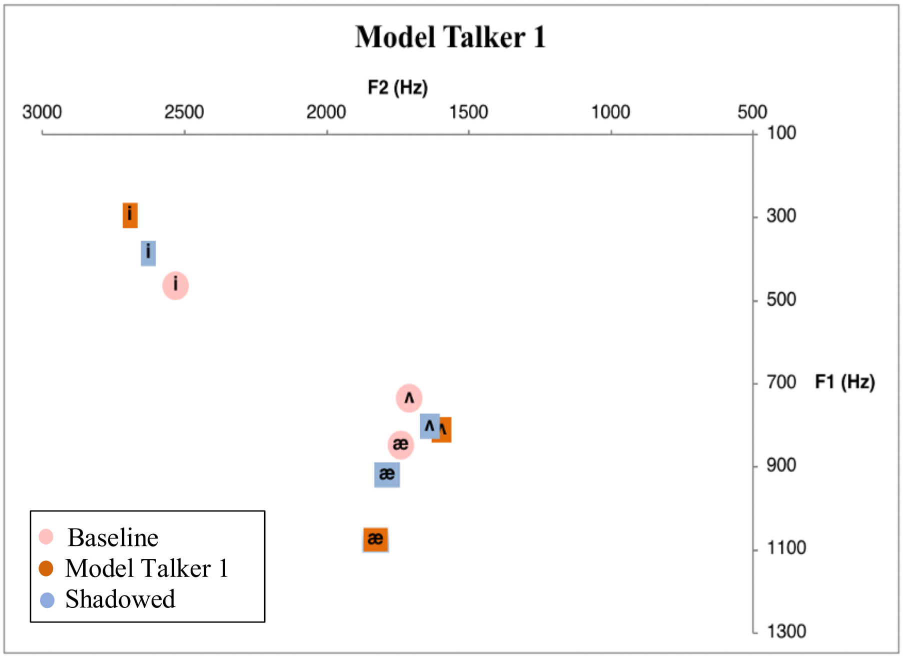


Figure 7. *Formant Plot Displaying Phonetic Convergence of All Learners to Model Talker 2*

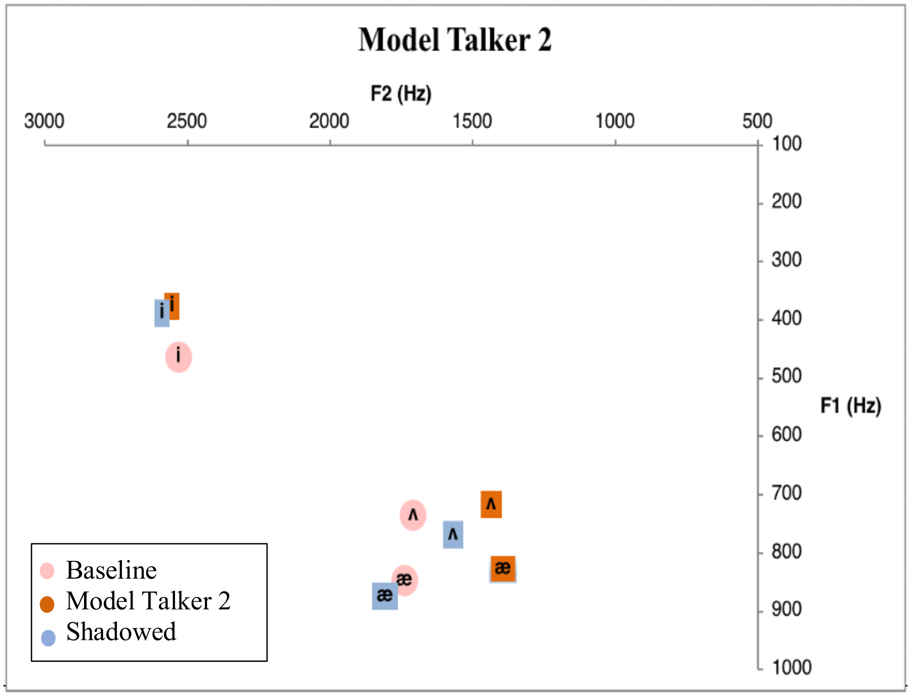


Figure 8. *Formant Plot Displaying Phonetic Convergence of All Learners to Model Talker 3*

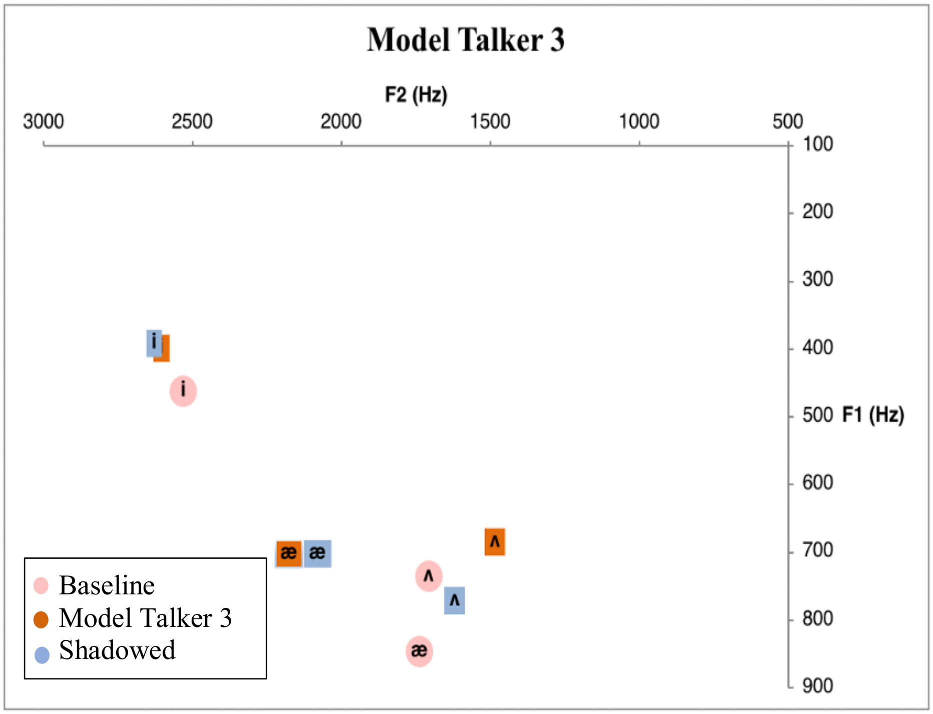


Figure 9. *Formant Plot Displaying Phonetic Convergence of All Learners to Model Talker 4*

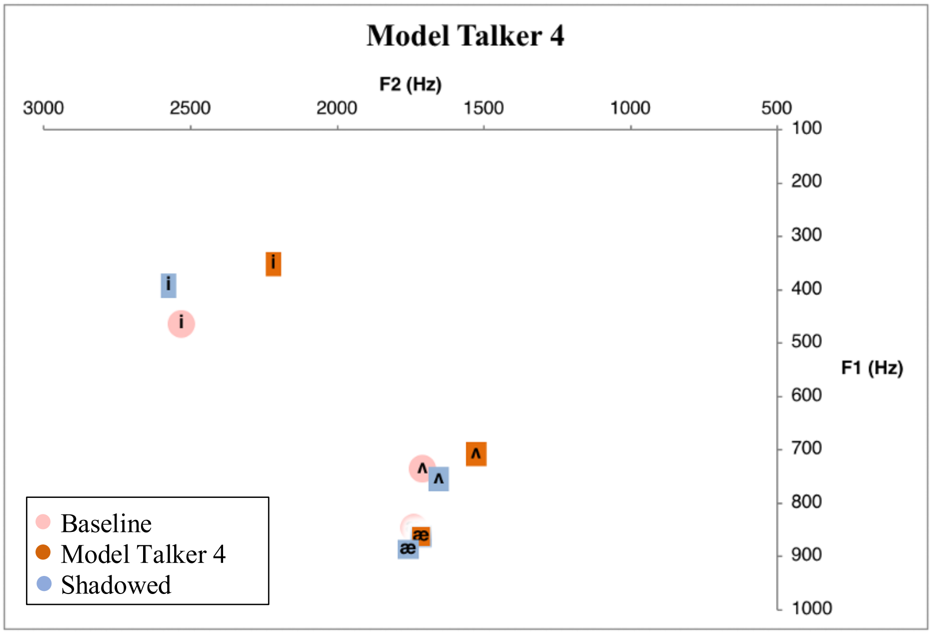
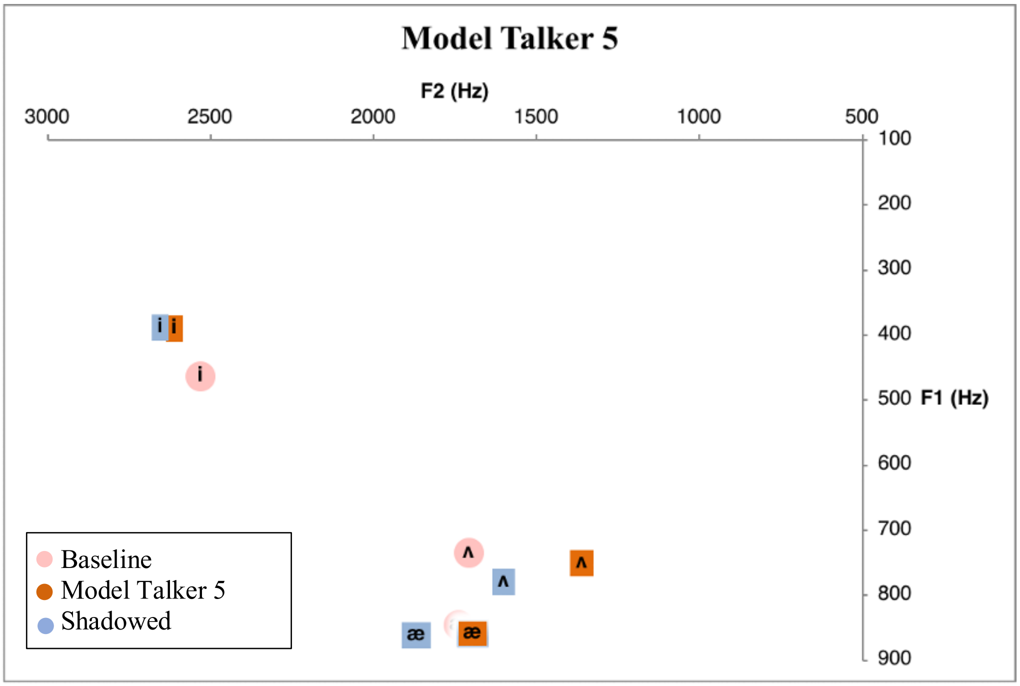


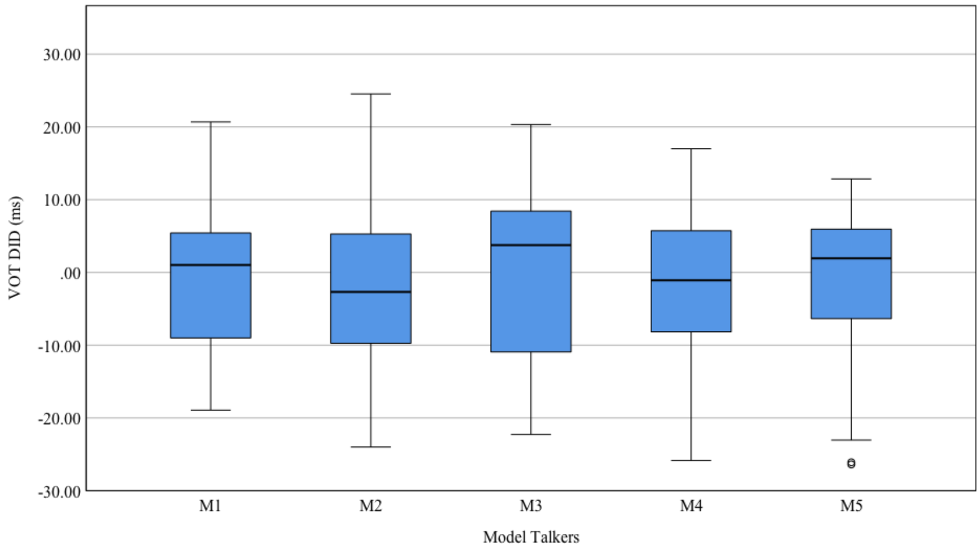
Figure 10. *Formant Plot Displaying Phonetic Convergence of All Learners to Model Talker 5*



3.7.2.1 VOT

Figure 11 describes the averaged VOT DID values of all stops of all learners with the five native model talkers. Positive values indicate convergence in the shadowing tasks, negative values indicate divergence, and zero indicates no change in vowel duration. Since all VOT DID values with the five model talkers were normally distributed, one-sample *t*-tests were used. They showed that all averaged VOT DIDs with the five model talkers were not significantly different from zero: Model Talker 1 (*M* = -.27 ms, *t*(25) = -.132, *p* > 0.05), Model Talker 2 (*M* = - 2.45 ms, *t*(25) = -.97, *p* > 0.05), Model Talker 3 (*M* = -.26 ms, *t*(25) = -.110, *p* > .05), Model Talker 4 (*M* = -1.13 ms, *t*(25) = -.551, *p* > 0.05), Model Talker 5 (*M* = -1.34 ms, *t*(25) = -.621, *p* > 0.05). These results generally suggested that participants maintained their VOT, or their VOT convergence did not reach a significant level.

Figure 11. *Boxplots of VOT DID Values in Milliseconds Illustrating the Degree and Direction of VOT Convergence towards the Five Model Talkers*



A one-way repeated measures ANOVA was conducted to compare VOT DID values across model talkers. The mean, standard deviation, median, and Kolmogorov-Smirnov test of normality are presented in Table 7. No significant effect was found for model talker: Wilks' Lambda = .95, $F(4, 22) = 301$, $p > .05$, multivariate partial eta squared = .05.

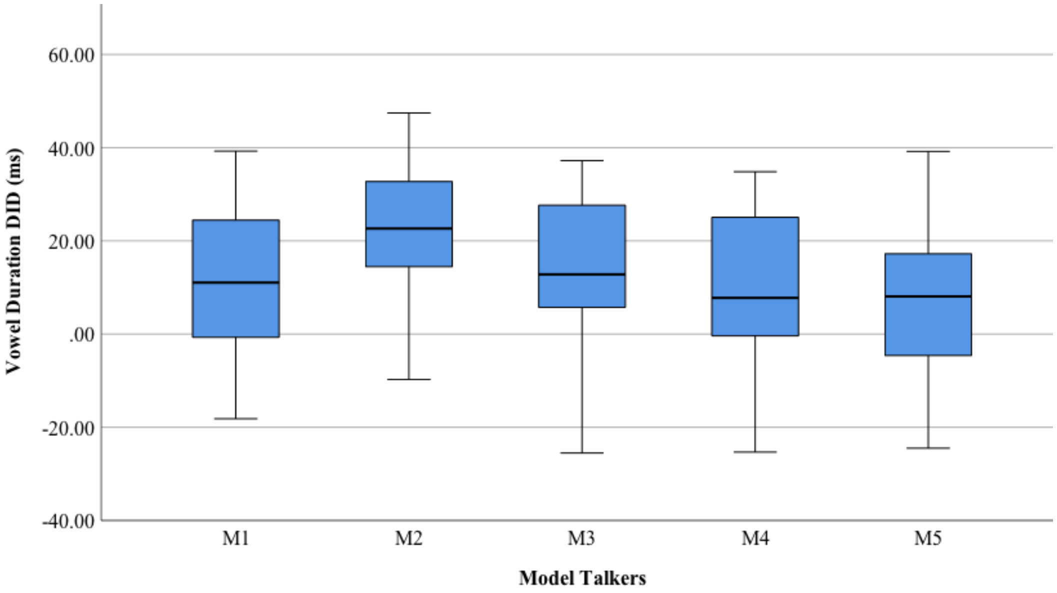
Table 7. *Descriptive Statistics of VOT DID Values across the Five Model Talkers*

Model Talker	*M*	*SD*	*Md*	*K-S*
1	-0.27	10.42	1.03	.200
2	-2.45	12.84	-2.67	.200
3	-0.26	12.04	3.76	.077
4	-1.13	10.44	-1.07	.200
5	-1.34	11.03	1.95	.143

3.7.2.2 Vowel Duration

Figure 12 presents the averaged vowel duration DID values of all vowels of all participants with the five model talkers. Positive values indicate convergence in the shadowing tasks, negative values indicate divergence, and zero indicates no change in vowel duration. First, normality checks were carried out on vowel duration DID values, showing these values were normally distributed across model talkers. Five one-sample *t*-tests revealed that vowel duration DID values in the five model talkers were significantly higher than zero. Participants showed the most vowel duration convergence towards Model Talker 2 (*M* = 21.36 ms, *SD* = 14.25, *t*(25) = 7.64, *p* = .000), followed by Model Talker 3 (*M* =14.10 ms, *SD* = 16.23, *t*(25) = 4.43, *p* = .000), Model Talker 1 (*M* = 10.38 ms, *SD* =15.82, *t*(25) = 3.35, *p* = .003), and Model Talker 4 (*M* = 9.56 ms, *SD* = 14.28, *t*(25) = 3.42, *p* = .002). However, less convergence was found towards Model Talker 5 (*M* = 7.10 ms, *SD* = 16.46, *t*(25) = 2.20, *p* = .03). These results indicated that learners converged their vowel duration towards all five model talkers.

Figure 12. *Boxplots of Vowel Duration DID Values in Milliseconds Illustrating the Degree and Direction of Vowel Duration Convergence towards the Five Model Talkers*



A one-way repeated measures ANOVA was conducted to examine whether vowel duration DID values significantly differed among model talkers. Table 8 presents the means and standard deviations of vowel duration DID values. A significant effect for model talker was found: Wilks' Lambda = .413, *F*(4, 22) = 7.82, *p* < .001, multivariate partial eta squared = .59. Pairwise comparisons with the Bonferroni correction revealed the DID values of Model Talker 2 (*M* = 21.36, *SD* = 14.25) were significantly higher than Model Talker 1 (*M* = 10.38, *SD* = 15.82, *p* =. 002), Model Talker 4 (*M* = 9.56, *SD* = 14.28, *p* =.000), and Model Talker 5 (*M* = 7.10, *SD* = 15.46, *p* = .000). These results suggested learners all converged their vowel duration to native English speakers but to different degrees.

Table 8. *Descriptive Statistics of Vowel Duration DID Values across the Five Model Talkers*

Model Talker	*M*	*SD*	*Md*	*K-S*
1	10.38	15.82	11.07	.200
2	21.36	14.25	22.68	.200
3	14.10	16.23	12.83	.200
4	9.56	14.28	7.79	.200
5	7.10	16.46	8.10	.200

3.7.2.3 F0

The averaged F0 DID values of all vowels of all learners with the five model talkers are shown in Figure 13. Positive values indicate convergence, negative values indicate divergence, and zero indicates no change. A normality check revealed that F0 DID values of Model Talkers 1, 2, 3, and 5 were normally distributed, but Model Talker 4 was not normally distributed. One-sample *t*-tests showed that DID values were not significantly different from zero for Model Talker 1 (*M* = 3.82 Hz, *SD* = 12.75, *t*(25) = 1.53, *p* > 0.05), Model Talker 2 (*M* = 5.31 Hz, *SD* = 16.60, *t*(25) = 1.63, *p* > .05), Model Talker 3 (*M* = 2.54 Hz, *SD* = 2.54, *t*(25) = .81, *p* > 0.05), or Model Talker 5 (*M* = -1.25, *SD* = 17.18, *t*(25) = -.370, *p* > .05). With Model Talker 4, a one-sample Wilcoxon signed rank test showed no significant difference from zero either (*Md* = -.67, *Z* = 1.00, *p* > .05). The average F0 DID values in Model Talkers 1, 2, and 3 were positive, indicating convergence, but not to a significant level. Table 9 summarizes the descriptive statistics of F0 DID values across the five model talkers.

Figure 13. *Boxplots of F0 DID Values in Milliseconds Illustrating the Degree and Direction of F0 Convergence towards the Five Model Talkers*

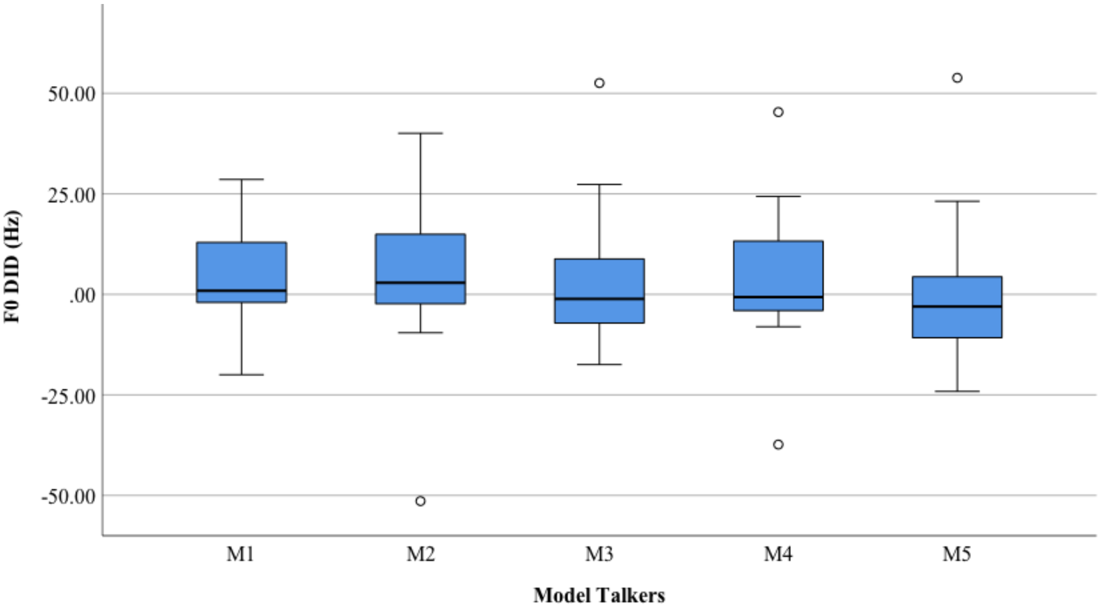


Table 9. *Descriptive Statistics of F0 DID Values across the Five Model Talkers*

Model Talker	*M*	*SD*	*Md*	*K-S*
1	3.82	12.75	.94	.17
2	5.31	16.60	2.90	.15
3	2.54	15.10	-1.10	.20
4	3.61	15.25	-.67	.02*
5	-1.25	17.18	-3.02	.19

Since not all F0 DID values were normally distributed, a one-way repeated measures ANOVA could not be used. An alternative non-parametric test, Friedman's two-way analysis of variance by ranks, was computed to test for differences in F0 DID values across model talkers. This test compares the difference in ranks using the median and thus is not influenced by the shape of the distribution. There was no significant difference in median F0 DID values across model talkers (χ^2 (4, *n* = 26) = 6.06, *p* > .05). Thus, learners generally did not show F0 convergence to model talkers, or convergence did not reach a significant level.

3.7.2.4 F1*F2

Figure 14 describes the averaged F1*F2 DID values of all vowels of all learners with the five model talkers. Positive values indicate convergence, negative values indicate divergence, and zero indicates no change in vowel duration. Tests of normality showed that F1*F2 DID values for Model Talkers 1 and 5 were not normally distributed, and a one-sample Wilcoxon signed rank test showed that the median F1*F1 DID values were not significantly different from zero for Model Talker 1 (*Md* = 30.05 Hz, *Z* = 1.66, *p* > .05) and Model Talker 5 (*Md* = -12.50 Hz, *Z* = -.93, *p* > .05). One-sample *t*-tests similarly showed that the averaged F1*F2 DID values were not significantly different from zero for Model Talker 2 (*M* = 23.27 Hz, *t*(25) = 1.16, *p* > 0.05), Model Talker 3 (*M* = 34.04 Hz, *t*(25) = 1.80, *p* > 0.05), or Model Talker 4 (*M* = 10.81 Hz, *t*(25) = .739, *p* > 0.05). Although the average and median F1*F2 DID values for Model Talkers 1, 2, 3, and 4 were positive, indicating convergence, the degree of convergence did not reach a significant level at 0.05. The average and median for Model Talker 5 were negative, indicating divergence, but not to a significant level.

The descriptive statistics of F1*F2 DID values across the five model talkers are displayed in Table 10. Since not all F1*F2 DID values were normally distributed, Friedman's two-way analysis of variance by ranks was conducted to examine if there was a difference in F1*F2 DID values across model talkers. Results revealed no significant difference in median F1*F2 DID values across the model talkers (χ^2 (4, *n* = 26) = 5.32, *p* > .05). In sum, the results generally suggested that Arabic-speaking English learners maintained their F1*F2, or convergence did not reach a significant level.

Figure 14. *Boxplots of F1*F2 DID Values in Hertz Illustrating the Degree and Direction of F1*F2 Convergence towards the Five Model Talkers*

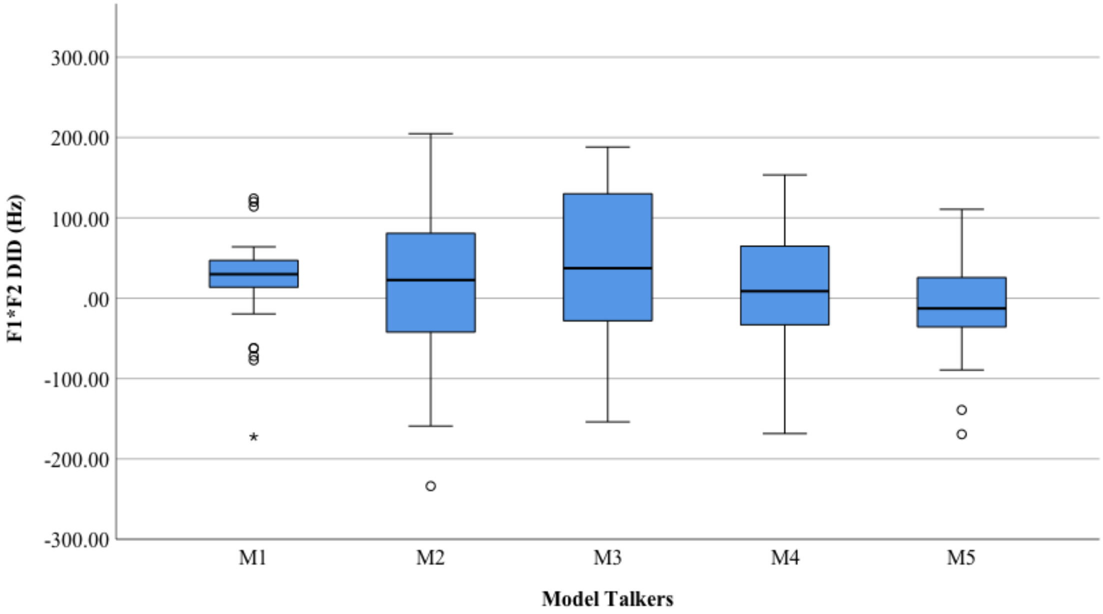


Table 10. *Descriptive Statistics of F1*F2 DID Values across the Five Model Talkers*

Model Talker	*M*	*SD*	*Md*	*K-S*
1	19.70	65.61	30.05	.001*
2	23.27	102.49	22.69	.182
3	34.04	96.26	37.52	.200
4	10.81	74.59	8.89	.200
5	-12.11	62.27	-12.50	.017*

3.8 Preexisting Phonetic Distance Results

To examine whether the preexisting phonetic distance between model talkers and participants was significantly associated with degree of convergence, the study employed GEE models with an identity link function. These models were appropriate for two reasons. First, the main aim was to find the relationship between preexisting phonetic distance and degree of convergence in general, not to a specific native model talker. Therefore, at every phonetic feature level, phonetic convergence data from learners in the five blocks were combined to ease data analysis. At a single phonetic dimension, each learner had five data points in each variable. The independent variable consisted of the absolute baseline distances between a given learner and the five model talkers. The dependent variable consisted of the DID values. All of the data were collapsed into four data subsets (VOT, vowel duration, F0, and F1*F2), resulting in repeated measures within a learner. Another reason for using GEE models was that most baseline preexisting distances were not normally distributed. Four models were used because the acoustic measurements were obtained from different scales. VOT and vowel duration were obtained in milliseconds, whereas F0, F1, and F2 were measured in Hertz.

The results of the four GEE models are summarized in Table 11. The B coefficient demonstrates the direction of the association between the dependent and independent variables. The same trend was observed in all of the examined acoustic features. The four absolute baseline distances (i.e., preexisting distances) had positive B coefficients, indicating the larger the baseline distance between learners and model talkers was, the larger the DID values would be.

The first GEE model found a significant relationship between VOT preexisting distance and learner DID values (B = .338, *p* = .014), suggesting this baseline was a significant predictor of DID. This relationship is demonstrated in Figure 15, where positive values indicate convergence in the shadowing tasks, negative values indicate divergence, and zero indicates no change. The absolute preexisting distances between learners and model talkers are on the x-axis. The degree of phonetic convergence in VOT (y-axis) was determined by the preexisting distances between learners and model talkers. This suggested the greater the preexisting distance, the larger the degree of VOT convergence was.

Table 11. *GEE Results for the Effect of Baseline Distance on DID Values*

			95% Wald Confidence Interval		Hypothesis Test	
Parameter	B	Std. Error	Lower	Upper	Wald Chi-Square	Sig.
Intercept	-6.019	1.93	-9.80	-2.24	9.75	.002
VOT	.338	.14	.07	.61	6.01	.014*
Intercept	-8.513	2.25	-12.91	-4.11	14.37	.000
V-duration	.814	.07	.70	.92	209.13	.000*
Intercept	-8.558	2.64	-13.73	-3.39	10.52	.001
F0	.497	.14	.22	.78	12.05	.001*
Intercept	-33.013	4.75	-42.33	-23.70	48.26	.000
F1*F2	.813	.09	.64	.99	83.10	.000*

Note. Dependent variable: DID values, independent variable: absolute baseline distance, variables at $p < 0.05$ are marked by an asterisk.

The second GEE model found that preexisting distance in vowel duration was a significant predictor of vowel duration DID values (B = .814, $p < .001$). Figure 16 illustrates this relationship, where positive values indicate convergence, negative values indicate divergence, and zero indicates no change. DID values on the y-axis show the degree of convergence in vowel duration, while the absolute preexisting distances between learners and model talkers are on the x-axis. Learners' degree of phonetic convergence in vowel duration was predicted by the magnitude of the preexisting distance between them and the model talkers. Thus, the greater the preexisting distance was, the more learners converged to the model talkers.

Figure 15. *Preexisting Distance and VOT DID Values*

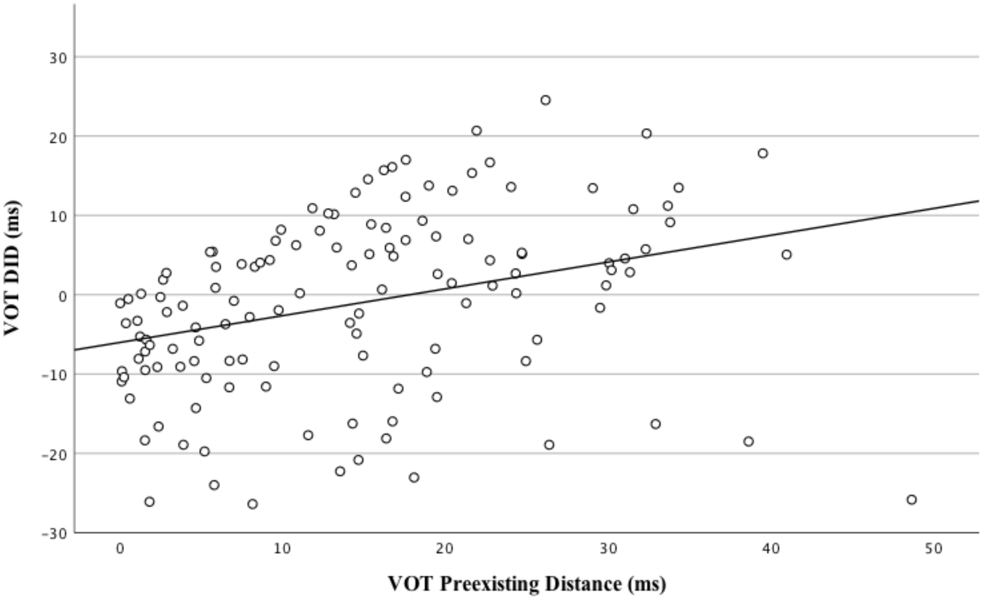
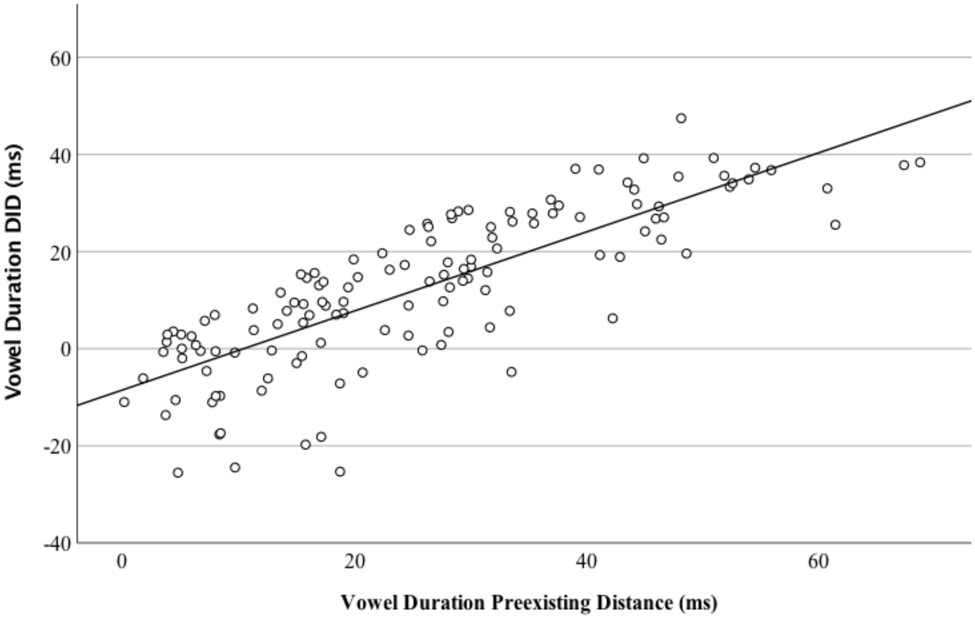
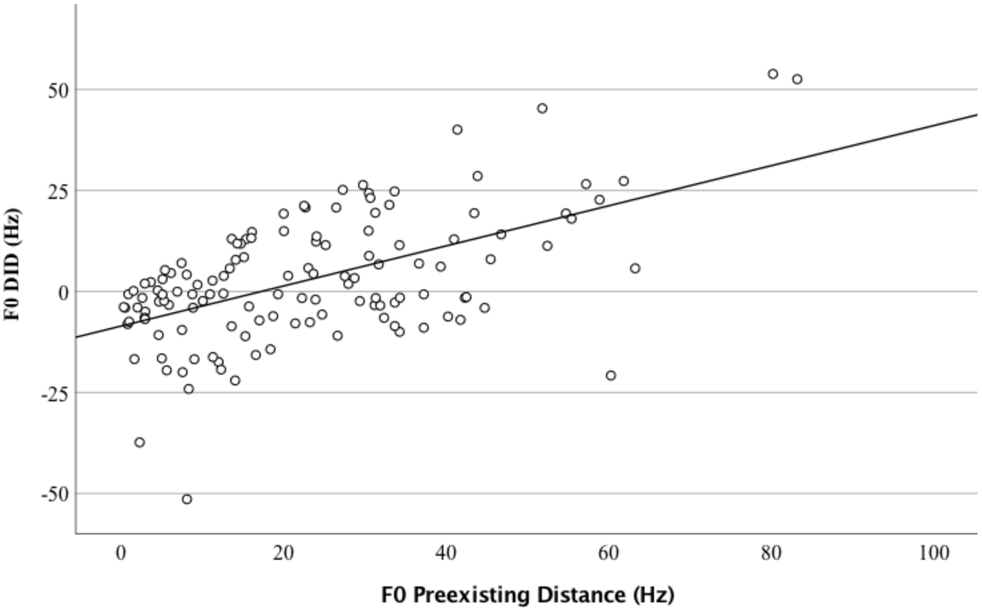


Figure 16. *Preexisting Distance and Vowel Duration DID Values*



In terms of F0, the GEE results again indicated preexisting distance was significantly associated with DID values (B = .497, *p* = .001). This relationship is illustrated in Figure 17, where positive values indicate convergence, negative values indicate divergence, and zero indicates no change. DID values on the y-axis show the degree of convergence in F0, while absolute preexisting distances between participants and model talkers are on the x-axis. These results indicated that with each Hz increase in the preexisting distance in F0, there was an associated increase in average DID values. In other words, the greater the preexisting distance was, the more participants converged to model talkers. This means that the degree of phonetic convergence in F0 was also determined by the preexisting distance between Arabic participants and model talkers.

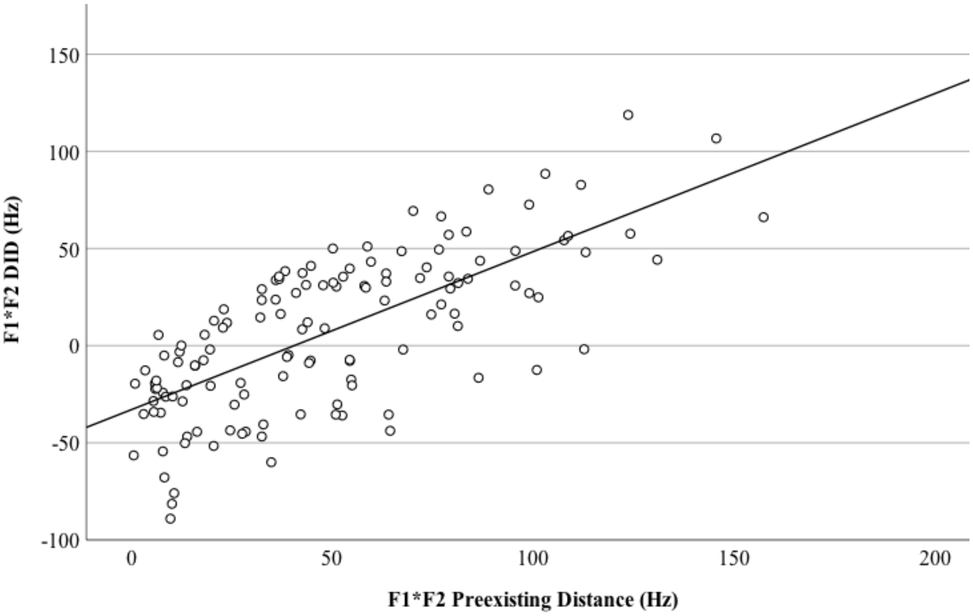
Figure 17. *Preexisting Distance and F0 DID Values*



The last GEE model showed preexisting distance was a significant predictor of F2*F2 DID values (B = .813, *p* < .001), and the scatter plot in Figure 18 shows a linear relationship between them. Positive values indicate convergence, negative values indicate divergence, and zero indicates no change. DID values on the y-axis show the degree of convergence in F1*F2, while the absolute preexisting distances between participants and model talkers are on the x-axis. The degree of phonetic convergence in F1*F2 was controlled by the preexisting distance between participants and model talkers.

Overall, the four GEE models revealed a direct relationship between the baseline distance and degree of convergence in all acoustic dimensions. That is, as the preexisting distance in a given phonetic dimension increased, the degree of convergence increased. These findings suggest learners showed more convergence to model talkers at a given phonetic dimension when their preexisting distance was further from the model talkers. Thus, the greater the baseline distance was, the larger the magnitude of phonetic convergence, as participants had more room to converge.

Figure 18. *Preexisting Distance and F1*F2 DID Values*



3.9 Relationship between Acoustic Measures and Perceived Phonetic Convergence

The aim of this section is to assess the impact of the acoustic DID values on judges' evaluation of phonetic convergence. Specifically, each acoustic DID measure was compared to the results obtained from the XAB perceptual tasks to find out which acoustic DID values predict listener judgments on phonetic convergence. To do so, GEE modeling was used to find the association between the acoustic measures and the perceived phonetic convergence. GEEs were used for the same reasons mentioned in the previous section. The analyses were performed on data collapsed across the five model talkers. Four GEE models with an independence correlation structure were built. Each acoustic feature was examined separately because DIDs were measured in different scales, as mentioned earlier - some were measured in hertz, and others were measured in milliseconds. The dependent variable in all the GEE models was the same, that is the average perceived phonetic convergence collapsed across the five perceptual similarities judgments (XAB). For each acoustic dimension, the independent variable was the average DID values of the 26 L2 learners, combined across the five model talkers. Thus, each learner had five data points in each GEE model.

Table 12 summarized the results from all generalized estimating equations models. The first GEE examined whether VOT DID had a contribution in the perceived phonetic convergence. Results indicated that there was a statistically significant relationship between the VOT DID values and the perceived phonetic convergence. VOT DID values were a significant predictor of listener judgments (B = .001, *p* = .039).

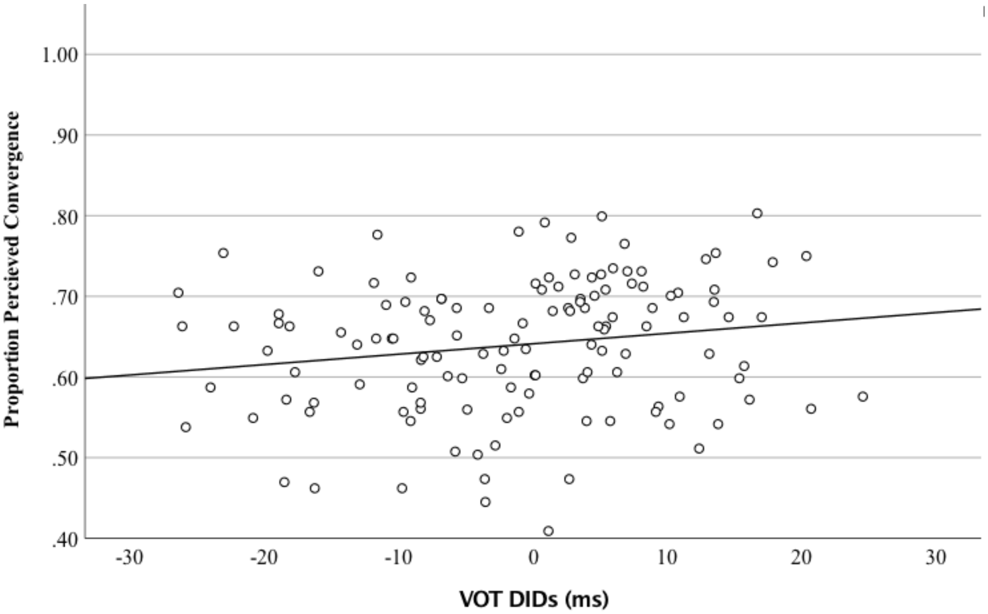
Table 12. *GEE Results for the Effect of DID Values on Perceptual Similarity Judgments (XAB)*

			95% Wald Confidence Interval			
Parameter	B	Std. Error	Lower	Upper	Wald Chi-Square	Sig.
(Intercept)	.64	.0076	.626	.656	7034.309	.000
VOT DID	.001	.0006	6.23	.003	4.241	.039*
(Intercept)	.631	.0116	.608	.654	2952.878	.000
V-duration DID	.001	.0005	.000	.002	2.414	.120
(Intercept)	.637	.0080	.621	.653	6304.624	.000
F0DID	.001	.0005	.000	.002	4.986	.026*
(Intercept)	.636	.0082	.619	.652	5948.084	.000
F1*F2 DID	.001	.0002	.000	.001	13.701	.000*

Note. Dependent variable: listeners' perceptual similarity judgments (XAB), independent variable: DID values, variables at $p < 0.05$ are marked by an asterisk.

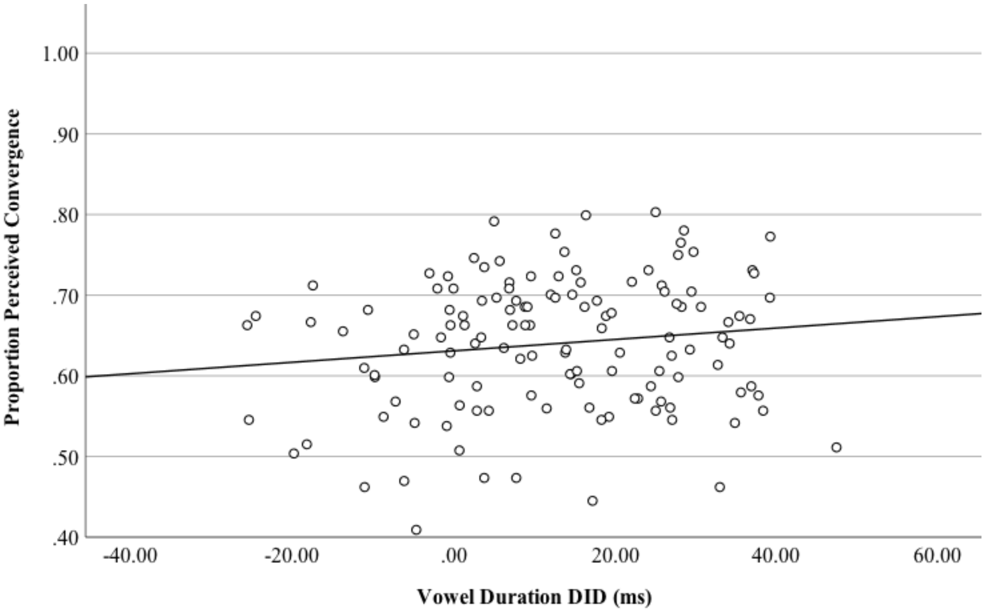
These data are plotted in Figure 19, where positive values indicate convergence, negative values indicate divergence, and zero indicates no change. The proportion of perceived phonetic convergence from the five XAB perceptual similarity tasks is on the y-axis, while VOT DID values are on the x-axis. Higher VOT DID values resulted in a higher probability that a native English listener would choose the shadowed productions as more similar to the model talkers than the baseline.

Figure 19. *VOT DID Values and Proportion of Perceived Phonetic Convergence*



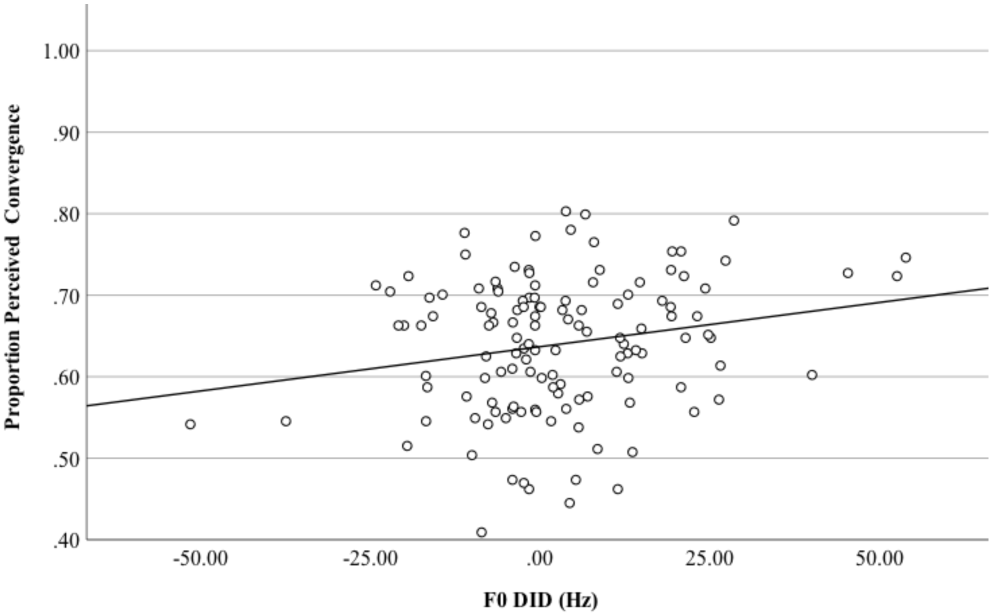
The second GEE model found no significant relationship between vowel duration DID values and perceived phonetic convergence (B = .001, *p* = .120). This indicated that vowel duration DID values were not a predictor of listener judgments, as illustrated in Figure 20, where positive values indicate convergence, negative values indicate divergence, and zero indicates no change. The proportion of perceived phonetic convergence from the five XAB perceptual similarity tasks is on the y-axis, and the vowel duration DID values are on the x-axis. Higher vowel duration DID values did not affect how listeners assessed phonetic convergence, although Arabic participants showed more convergence with vowel duration than with the other phonetic attributes. Thus, listeners did not build their judgements on vowel duration convergence.

Figure 20. *Vowel Duration DID Values and Proportion of Perceived Phonetic Convergence*



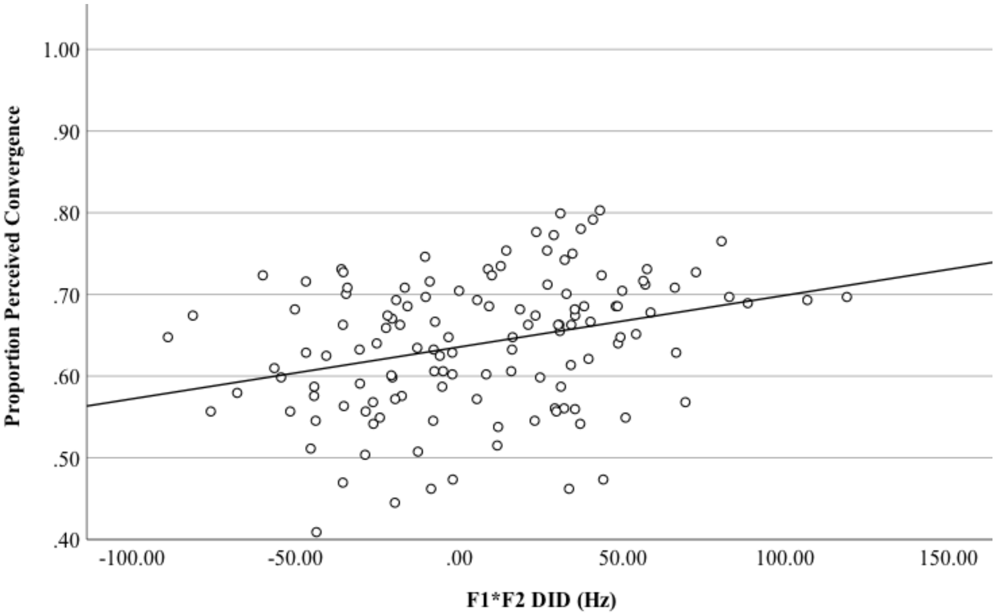
The third GEE model found a significant relationship between F0 DID values and perceived phonetic convergence, indicating F0 DID values were a significant predictor of listener judgments (B = .001, *p* = .026). These data are illustrated in Figure 21 by a scatter plot. The proportion of perceived phonetic convergence from the five XAB perceptual similarity tasks is on the y-axis, and the F0 DID values are on the x-axis. Positive values indicate convergence, negative values indicate divergence, and zero indicates no change. The probability of selecting shadowed items as more similar to model talkers increased as the F0 DID values increased.

Figure 21. *F0 DID Values and Proportion of Perceived Phonetic Convergence*



The last GEE model revealed a significant relationship between F1*F2 DID values and perceived phonetic convergence, suggesting F1*F2 DID values were a significant predictor of native English listeners' judgments (B = .001, *p* = .000). In other words, the higher the F1*F2 DID values were, the more listeners selected the shadowed items as more similar to the model talkers than to the baseline. Figure 22 represents the direct relationship between these two variables. The proportion of perceived phonetic convergence from the five XAB perceptual similarity tasks is on the y-axis, and the F1*F2 DID values are on the x-axis. Positive values indicate convergence in the shadowing tasks, negative values indicate divergence, and zero indicates no change.

Figure 22. *DID Values and Proportion of Perceived Phonetic Convergence*



The GEE models suggested that VOT, F0, and F1*F2 DID values had a direct relationship to perceived phonetic convergence. Thus, the higher these values were, the more likely native English listeners were to select the shadowed items as similar to model talkers. In other words, the more learners converged their VOT, F0, and F1*F2 to model talkers, the more they were perceived by native English listeners as similar to native speakers. Although statistical analysis showed that VOT, F0, and F1*F2 DID values were not significantly different from zero, native English listeners were sensitive to changes in these acoustic attributes. Vowel duration DID values were the only phonetic dimension that reached a significant level of convergence, with learners converging vowel duration more than any other acoustic attribute. Nevertheless, vowel duration similarities were not a significant predictor of listener judgments.

CHAPTER 4: PHONETIC PRODUCTION TRAINING

4.1 Phonetic Production Training

This chapter discusses the phonetic production training. It first gives an overview of how the phonetic convergence results were used to match trainees with their trainers and how they were divided into the convergence group (C-group) and divergence group (D-group). It then explains how the training sessions were designed, including a detailed description of the stimuli and procedures. After that, it describes the native-speaker intelligibility assessments of each group's performance. Finally, the chapter examines the results of the training sessions based on the intelligibility assessments.

4.2 Matching Trainers with Trainees

The trainees consisted of 20 of the 26 Saudi women who participated in the phonetic convergence experiment. The age of C-group trainees ranged from 25 to 47 (*M* = 31.1, *SD* = 6.33), while D-group ranged from 20 to 33 (*M* = 25.6, *SD* = 4.43). The five model talkers from the first experiment were also the trainers for the training sessions, and the results of the phonetic convergence experiment were used to match trainees with their trainers. Since each trainee would have only one trainer in the sessions, the following criteria were established. Perceptual similarity judgments were considered first, as they were all-inclusive since listeners had access to many acoustic phonetic properties of participants' productions. Additionally, these judgements minimized the misrepresentative explanations that can be found in analyzing just one acoustic feature (Prado at el., 2017). I followed these criteria since phonetic convergence is complex, and it is difficult to predict the specific degrees or direction of convergence in all attributes.

First, the highest and lowest perceived convergence ratios by model talkers were examined for each learner to divide the remaining 20 into two even groups. Table 13 compares these scores in rank order from lowest to highest convergence. No data are listed for Participant 10 as she did not provide complete data. Participant numbers were kept the same between the first experiment and the training to ease the analysis and avoid confusion. When selecting numeric values for each group, I first considered the proportion of perceived phonetic convergence, as mentioned above. However, when multiple model talkers showed similar proportions of perceived phonetic convergence to a participant, I selected a single model talker based on comparing DID values within the same value but not across different measures. Ten of the 20 trainees received similar or relative proportions of perceived convergence in either direction (Participants 1, 3, 8, 13, 16, 20, 23, 24, 26, and 27).

Table 13. *Mean Proportion of Perceived Phonetic Convergence towards the Models Talkers in Ascending Order for Each Trainee*

Participant	Lowest Convergence				Highest Convergence
1	M5 (0.50)	M2 (0.60)	M3 (0.66)	M1 (0.67)	M4 (0.68)
2	M2 (0.46)	M4 (0.65)	M3 (0.67)	M5 (0.69)	M1 (0.75)
3	M2 (0.62)	M4 (0.63)	M3 (0.70)	M5 (0.70)	M1 (0.73)
6	M4 (0.57)	M2 (0.59)	M1 (0.64)	M5 (0.72)	M3 (0.78)
8	M3 (0.55)	M5 (0.58)	M1 (0.63)	M4 (0.65)	M2 (0.67)
9	M4 (0.54)	M2 (0.56)	M1 (0.65)	M3 (0.67)	M5 (0.70)
11	M4 (0.56)	M2 (0.65)	M1 (0.66)	M5 (0.68)	M3 (0.80)
13	M1 (0.62)	M2 (0.62)	M4 (0.63)	M3 (0.66)	M5 (0.66)
14	M2 (0.57)	M4 (0.60)	M5 (0.69)	M1 (0.71)	M3 (0.73)
15	M2 (0.58)	M5 (0.64)	M4 (0.67)	M3 (0.73)	M1 (0.77)
16	M4 (0.47)	M1 (0.63)	M2 (0.69)	M3 (0.71)	M5 (0.71)
18	M5 (0.41)	M1 (0.46)	M4 (0.47)	M2 (0.59)	M3 (0.66)
20	M2 (0.57)	M5 (0.60)	M1 (0.65)	M3 (0.66)	M4 (0.67)
21	M2 (0.61)	M4 (0.63)	M3 (0.68)	M5 (0.70)	M1 (0.73)
22	M2 (0.55)	M4 (0.66)	M5 (0.69)	M1 (0.69)	M3 (0.72)
23	M1 (0.55)	M4 (0.56)	M2 (0.56)	M5 (0.68)	M3 (0.70)
24	M4 (0.56)	M1 (0.56)	M2 (0.57)	M3 (0.70)	M5 (0.73)
25	M5 (0.56)	M2 (0.66)	M4 (0.71)	M1 (0.72)	M3 (0.74)
26	M1 (0.60)	M2 (0.63)	M4 (0.69)	M3 (0.70)	M5 (0.71)
27	M4 (0.51)	M2 (0.60)	M5 (0.61)	M3 (0.69)	M1 (0.72)

Note. M stands for model talker. The number in parentheses is the mean proportion of perceived convergence towards the model talkers.

To match these ten trainees with their trainers, the phonetic convergence in each of the four acoustic attributes was evaluated separately in relation to the model talkers based on the DID values. Positive values indicated convergence, negative values indicated divergence, and zero or near-zero values indicated that participants neither converged nor diverged from model talkers, maintaining their acoustic values. No zero values were found in the DID values, but there were some near-zero values. For each model talker, the number of acoustic features that participants converged to was calculated. If one model talker exhibited more convergence in acoustic features, she was chosen as the model talker with the highest degree of convergence. If a participant had the same number of converged features towards multiple model talkers, the exact DID values were compared within each acoustic measure, not across different measures. The final decision on the model talker was made based on the degree of convergence that participants showed.

For example, Participant 1 received a 0.67 with Model Talker 1 and 0.68 with Model Talker 4 as the highest proportions of perceived convergence. The proportion with Model Talker 4 was greater, but the difference was small. Model Talker 4 was determined to have the highest proportion as she showed convergence in all four acoustic measures (VOT, vowel duration, F0, and F1*F2), while she diverged in all phonetic features from Model Talker 1.

Another example is Participant 13, who obtained a 0.66 with Model Talkers 3 and 5 in highest perceived phonetic convergence. Model Talker 3 was chosen because the participant converged all acoustic features to her but converged only VOT and F0 to Model Talker 5. At the same time, this participant received nearly the same perceptual assessment scores for Model Talkers 1, 2, and 4. To decide which she converged to the least, acoustic convergence was considered. She converged two acoustic features to each model talker, so the exact DID values were compared. The lowest degree of convergence was found towards Model Talker 1. Thus, it was concluded that this participant phonetically converged the least to her. The same criteria were followed with the eight remaining participants. After resolving similar issues, the highest and lowest scores and the associated model talkers were entered into a comparison table to divide the participants into two groups (see Table 14).

Table 14. *Highest and Lowest Convergence with Model Talkers in the Perceptual Similarity Tests*

Participant	Lowest Convergence	Highest Convergence
1	M5 (0.50)	M4 (0.68)
2	M2 (0.46)	M1 (0.75)
3	M4 (0.63)	M1 (0.73)
6	M4 (0.57)	M3 (0.78)
8	M3 (0.55)	M2 (0.67)
9	M4 (0.54)	M5 (0.70)
11	M4 (0.56)	M3 (0.80)
13	M1 (0.62)	M3 (0.66)
14	M2 (0.57)	M3 (0.73)
15	M2 (0.58)	M1 (0.77)
16	M4 (0.47)	M3 (0.71)
18	M5 (0.41)	M3 (0.66)
20	M2 (0.57)	M3 (0.66)
21	M2 (0.61)	M1 (0.73)
22	M2 (0.55)	M3 (0.72)
23	M1 (0.55)	M5 (0.68)
24	M4 (0.56)	M5 (0.73)
25	M5 (0.56)	M3 (0.74)
26	M1 (0.60)	M3 (0.70)
27	M4 (0.51)	M3 (0.69)

The convergence group (C-group) was trained by the model talkers to whom they showed the most convergence, whereas the divergence group (D-group) was trained by the model talkers to whom they showed divergence or the least convergence. To decide which group each participant would be assigned to, the highest proportions were listed in descending order for C-group, as represented in Table 15, and the lowest proportions were listed in ascending order for D-group, as shown in Table 16. Within each table, the first 10 participants were chosen to be assigned to the two training groups.

Table 15. *Highest Convergence in the Perceptual Similarity Tests (XAB) in Descending Order*

Participant	Highest perceived convergence	Model Talker
11	0.8	3
6	0.78	3
15	0.77	1
2	0.75	1
25	0.74	3
3	0.73	1
14	0.73	3
21	0.73	1
24	0.73	5
22	0.72	3
16	0.71	3
9	0.7	5
26	0.7	3
27	0.69	3
1	0.68	4
23	0.68	5
8	0.67	2
13	0.66	3
18	0.66	3
20	0.66	3

When the same participant appeared in the 10 highest and lowest lists simultaneously, her ranking in both lists was considered to assign her to one of the groups. For example, Participant 11 was in the highest rank in the highest convergence list as well as the tenth lowest rank in the lowest convergence list. She was assigned to C-group as the first rank is higher than the tenth rank. This decision was made to maximize the degree of convergence. Another example is Participant 2, who had the fourth highest rank in the highest convergence list but the second lowest rank in the lowest convergence list. In this case, Participant 2 was assigned to D-group since the second rank is higher than the fourth. The same criteria were followed with the remaining participants.

Table 16. *Lowest Convergence in the Perceptual Similarity Tests (XAB) in Ascending Order*

Participant	Lowest perceived convergence	Model Talker
18	0.41	5
2	0.46	2
16	0.47	4
1	0.5	5
27	0.51	4
9	0.54	4
8	0.55	3
22	0.55	2
23	0.55	1
11	0.56	4
24	0.56	4
25	0.56	5
6	0.57	4
14	0.57	2
20	0.57	2
15	0.58	2
26	0.6	1
21	0.61	2
13	0.62	1
3	0.63	4

The final training groups are shown in Tables 17 and 18. Table 17 represents the trainees with their associated trainers in C-group. Six out of 10 received training from Model Talker 3, three received training from Model Talker 1, and only one was trained by Model Talker 5. In D-group, as illustrated in Table 18, three participants were trained by Model Talker 2 and three were trained by Model Talker 4. Model Talkers 1 and 3 had only one trainee each. The remaining two were trained by Model Talker 5, while Model Talker 2 did not train anyone in C-group. This was expected as she received the lowest perceived convergence during shadowing out of all model talkers (*M* = 0.59).

Table 17. *Trainees in C-Group with Their Associated Trainers*

Trainee	Trainer	Perceived Convergence
11	M3	0.80
6	M3	0.78
15	M1	0.77
25	M3	0.74
3	M1	0.73
14	M3	0.73
21	M1	0.73
24	M5	0.73
26	M3	0.7
13	M3	0.66

Note. The third column shows the proportion of perceived convergence in descending order.

Table 18. *Trainees in D-Group with Their Associated Trainers*

Trainee	Trainer	Perceived Convergence
18	M5	0.41
2	M2	0.46
16	M4	0.47
1	M5	0.5
27	M4	0.51
9	M4	0.54
8	M3	0.55
22	M2	0.55
23	M1	0.55
20	M2	0.57

Note. The third column shows the proportion of perceived convergence in ascending order.

4.3 Differences Between the Two Groups

4.3.1 Perceived Convergence

Normality checks on the perceived convergence scores for both groups revealed they were normally distributed, as shown in Table 19.

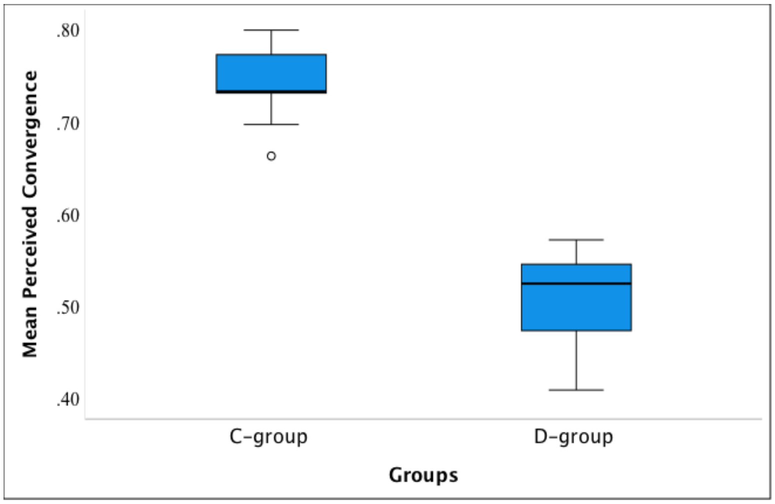
Table 19. *Descriptive Statistics of Perceived Convergence for C-Group and D-Group*

Group	*M*	*SD*	*Md*	*K-S*
C-group	.74	.04	.73	.138
D-group	.51	.05	.52	.146

Note. K-S < .05 means data were not normally distributed.

A two-tailed one-sample *t*-test compared C-group's proportions to chance level (0.5) and found them significantly higher (*M* = .74, *SD* = .04, *t*(9) = 19.113, *p* < .001). This means native English listeners chose the shadowed productions more often than the baseline, indicating phonetic convergence. Another two-tailed one-sample *t*-test found no significant difference between D-group's scores and chance level (*M* = .51, *SD* = .05, *t*(9) = .691, *p* > .05). In other words, listeners were not able to distinguish shadowed productions from the baseline, indicating a lack of convergence. Finally, an independent-samples *t*-test showed the mean score for C-group was significantly higher (*M* = .74, *SD* = .04) than D-group (*M* = .51, *SD* =.05), *t*(18) = 12.69, *p* <. 001, two-tailed. Figure 23 presents these results in box-plot format.

Figure 23. *Perceived Convergence Proportions for C-Group and D-Group*



4.3.2 Convergence in Acoustic Features

This section compares C-group and D-group in terms of DID values, which included VOT, vowel duration, F0, and F1*F2. The first step was to ensure all data were normally distributed before conducting parametric tests. A Kolmogorov-Smirnov test of normality revealed that all DID values for both groups were normally distributed (see Table 20).

Table 20. *Phonetic Convergence in Acoustic Features for C-Group and D-Group*

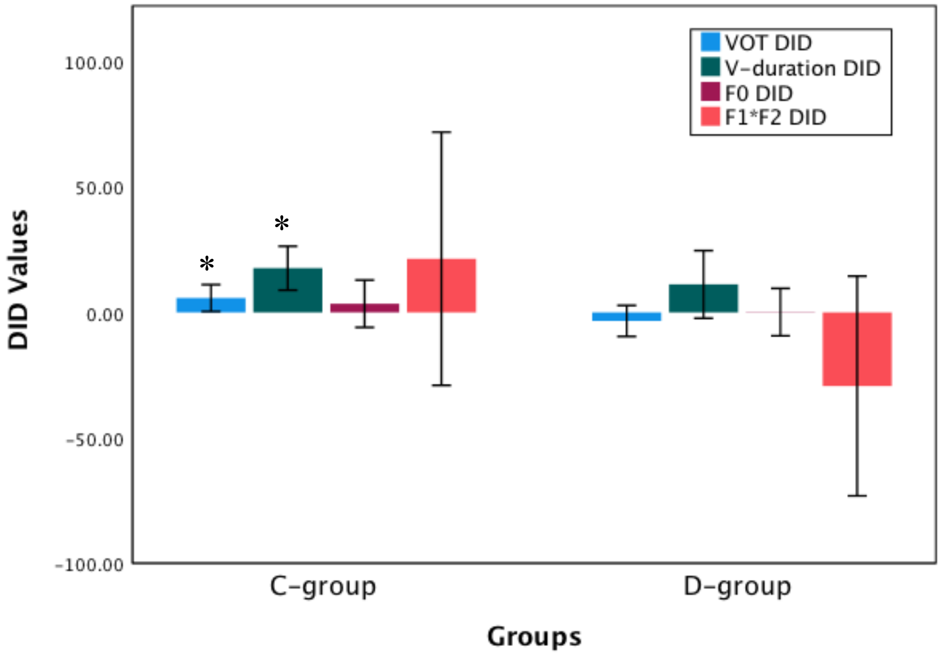
DID	Group	*M*	*SD*	*Md*	*K-S*
VOT	C-group	6.19	7.39	6.45	.119
	D-group	-3.37	8.66	-3.85	.925
V-duration	C-group	17.64	12.21	13.92	.124
	D-group	11.18	18.85	13.06	.609
F0	C-group	3.48	13.16	2.56	.921
	D-group	.14	13.20	-4.49	.355
F1*F2	C-group	21.41	70.56	17.60	.154
	D-group	-29.32	61.25	-33.30	.964

Note. K-S < .05 means the data were not normally distributed.

Within each group, the DID values of the four acoustic attributes were analyzed to determine if participants' phonetic convergence was significantly higher than zero. For C-group, a one-sample *t*-test showed that VOT DID (*M* = 6.19 ms, *SD* = 7.39, *t*(9) = 2.65, *p* = 0.03) and V-duration DID (*M* = 17.64 ms, *SD* = 12.21, *t*(9) = 4.569, *p* = .001) were significantly higher than zero, while F0 DID (*M* = 3.48 Hz, *SD* = 13.16, *t*(9) = .837, *p* > .05) and F1*F2 DID (*M* = 21.41 Hz, *SD* = 70.56, *t*(9) = .959, *p* > .05) were not.

For D-group, a one-sample *t*-test showed that all DID values were not significantly different from zero, including for VOT (*M* = -3.37 ms, *SD* = 8.66, *t*(9) = -1.232, *p* > .05), V-duration (*M* = 11.18 ms, *SD* = 18.85, *t*(9) = 1.875, *p* > .05), F0 (*M* = .14 Hz, *SD* = 13.20, *t*(9) = .034, *p* > .05), and F1*F2 (*M* = -29.32 Hz, *SD* = 61.25, *t*(9) = -1.514, *p* > .05). This suggested D-group on average did not converge to the model talkers to a significant level. Figure 24 demonstrates the DID values across the four acoustic measures in both groups.

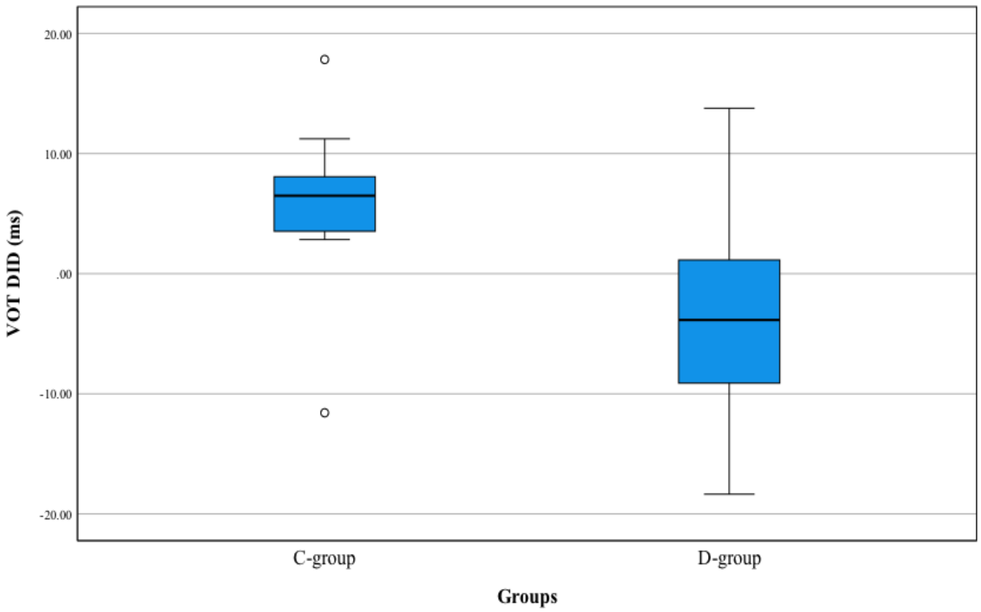
Figure 24. *Average DID Values across the Four Acoustic Measures in the Two Groups*



Note. Only VOT and V-duration in C-group were significantly higher than zero at the 0.05 level, marked by an asterisk. Error bars indicate standard errors. Positive values indicate convergence, negative values indicate divergence, and zero indicates no change.

To determine whether there was a significant difference between groups in terms of convergence of acoustic attributes, four independent-samples *t*-tests were conducted. Results showed that VOT DID values in C-group (*M* = 6.19 ms, *SD* = 7.39) were significantly higher (*t*(18) = 2.53, *p* = .02) than D-group (*M* = -3.37, *SD* = 8.66). Figure 25 illustrates these results in box-plot format.

Figure 25. *VOT DID Values in Milliseconds Illustrating the Degree and Direction of VOT Convergence towards the Trainers*



The statistical analysis revealed no significant difference in vowel duration convergence between C-group (*M* = 17.64, *SD* = 12.21) and D-group (*M* = 11.18, *SD* = 18.85), *t*(18) = .911, *p* > 0.05, two-tailed, nor was there a significant difference between C-group (*M* = 3.48, *SD* = 13.16) and D-group (*M* = 0.14, *SD* = 13.20) in terms of F0 (*t*(18) = .57, *p* > .05, two-tailed). Finally, there was no significant difference between groups in the convergence of F1*F2 (*t*(18) = 1.72, *p* > 0.05, two-tailed) despite C-group (*M* = 21.41, *SD* = 70.56) showing higher values than D-group (*M* = -29.32, *SD* = 61.25). These results are represented in Figures 26, 27, and 28, respectively. Positive values indicate convergence, negative values indicate divergence, and zero indicates no change.

Figure 26. *V-Duration DID Values in Milliseconds Illustrating the Degree and Direction of V-Duration Convergence towards the Trainers*

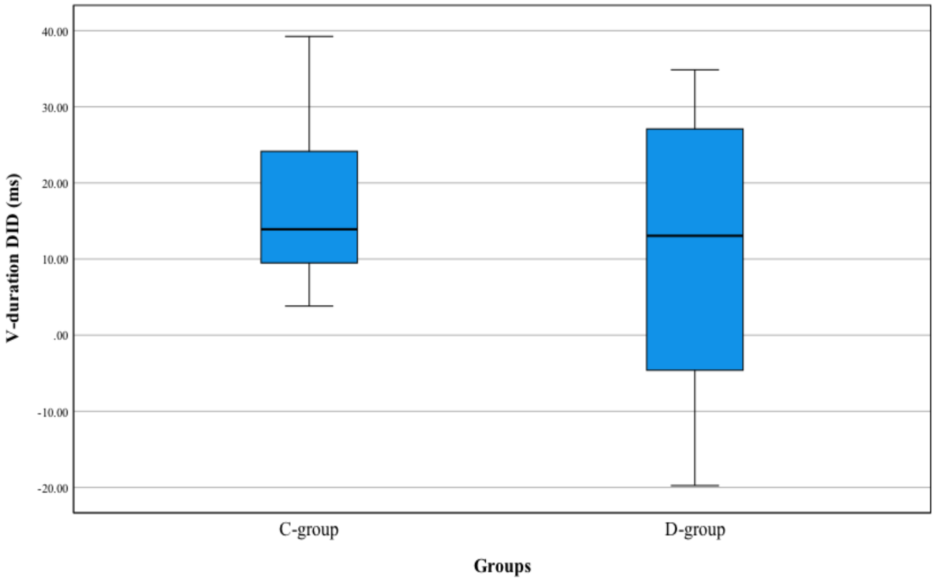


Figure 27. *F0 DID Values in Hertz Illustrating the Degree and Direction of F0 Convergence towards the Trainers*

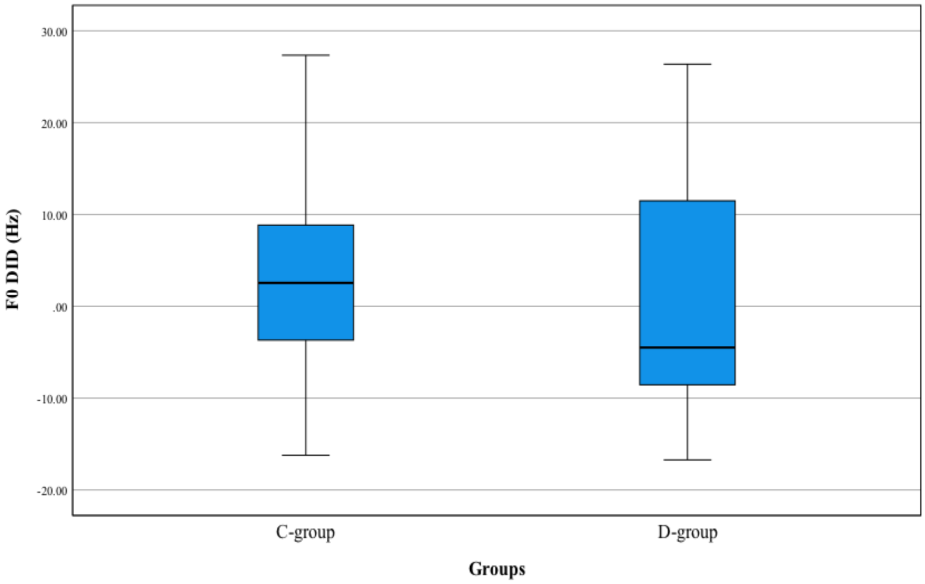
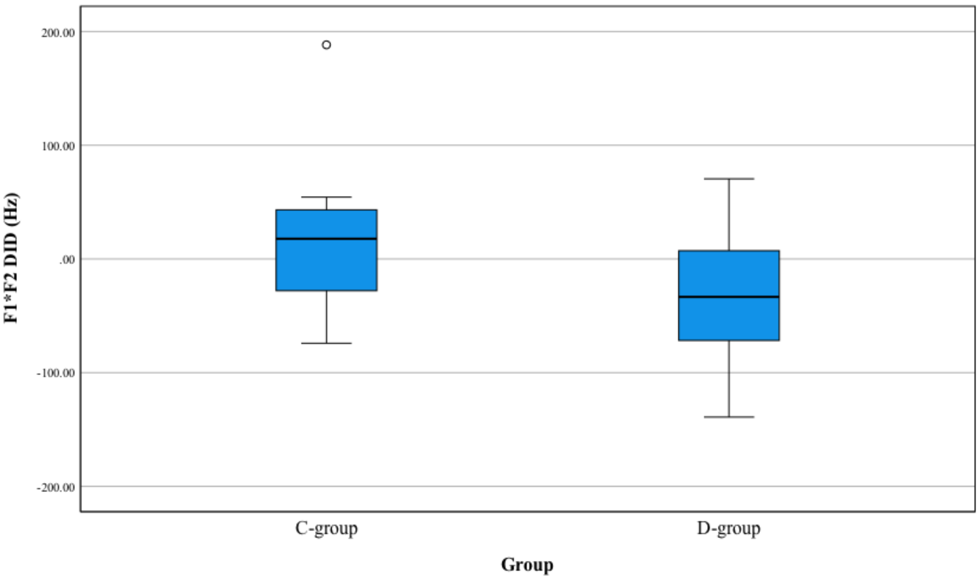


Figure 28. *F1*F2 Values in Hertz Illustrating the Degree and Direction of F1*F2 Convergence towards the Trainers*



The two groups were only significantly different from each other in terms of perceived convergence and VOT convergence, with C-group scoring higher than D-group. Nevertheless, there were obvious differences. For C-group, acoustic convergence in VOT and vowel duration (i.e., DID values) were significantly higher than zero. On average, C-group also tended to have greater convergence in all acoustic measures towards trainers in the training sessions. In contrast, D-group only showed convergence in vowel duration, though the effect was weaker than in C-group, and this convergence did not reach a significant level. VOT and F1*F2 DID on average showed divergence from trainers in D-group, while average F0 values witnessed no change. Thus, in addition to higher perceived convergence, C-group had higher average values for all acoustic measures than D-group as well, though only the difference in VOT convergence reached a significant level.

4.4 Training Sessions

4.4.1 Training Stimuli

A total of 12 nonsense words following English phonotactics were created to assess the effectiveness of the production training. The five native English speakers modeled in the phonetic convergence experiment produced these words. They were recorded individually in a quiet room using Audacity (Version 2.4.2) at the University of Wisconsin-Milwaukee. Soundproof booths were not used due to the pandemic at the time of the study. All spoken materials were recorded on a personal laptop connected to Focusrite Scarlett Solo Audio Interface (3rd Gen) through a UHURU mounted cardioid microphone with a sampling rate of 44,100 Hz. A microphone isolation shield was used to lessen background noise.

Following Barriuso (2018), the nonsense words were presented in sentences that had real rhyming words to help the model talkers pronounce them. For example, *duct* rhymes with *puct*, and *left* rhymes with *seft* (see Appendix A for the complete list). Sentences were presented to the model talkers one at a time in random order during the recording session to prevent listing intonation. The sentences were spoken by each model talker three times. For each model talker, the production that had the clearest pronunciation with the best auditory quality was selected for the training sessions.

The wordlist was divided into two sets: one containing target consonants and the other containing target vowels. To make the task more manageable and less overwhelming for the trainees, each set of training words had only one structure typically difficult for Arabic speakers learning English. The first set consisted of two subsets of monosyllabic words with word-initial consonants that Arabic speakers have problems with. The first subset contained three words beginning with the bilabial voiceless stop /p/, while the second contained three words beginning with the voiced interdental fricative /v/. All words in this set had a CVCC syllable structure, and the final consonant cluster had either falling or plateau sonority, as this structure was less likely to pose a difficulty for Saudi Arabic speakers. In particular, it is permissible in Najdi and Hijazi Arabic (Alfaifi, 2019), which were the native dialects of the majority of trainees.

The second set consisted of six monosyllabic words divided into two subsets. The first subset had the monophthong /ɛ/ in the nucleus of CVCC nonsense words. The second subset contained the diphthong /oʊ/ in the nucleus of CVC nonsense words. Again, no sounds or structures were expected to cause any difficulty in this set of words except the intended training vowels. Table 21 lists the complete set of nonsense words.

Table 21. *Nonsense Words Used in the Production Training Sessions*

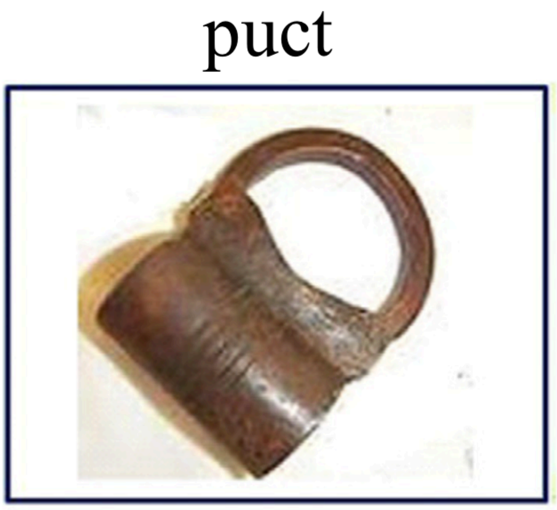
Set	Target segment	Nonsense word
Difficult consonants	Bilabial voiceless stop /p/	puct
		pusk
		paffs
	Voiced interdental fricative /v/	vant
		vulb
		vilt
Difficult vowels	Monophthong /ɛ/	ceft
		mest
		lesk
	Diphthong /oʊ/	roke
		sote
		fote

4.4.2 Training Procedure

All trainees were trained and tested according to the same procedure: a pre-test, three sessions of low-variability phonetic production training, and a post-test. During the pre-test and post-test phases, audio recordings were made of the trainees producing the English nonsense words, which were presented in writing along with an image of a nonobject on a screen. The images were taken from Bürki et al. (2012) and are given in Appendix B. The recording lasted about 10 minutes for each test.

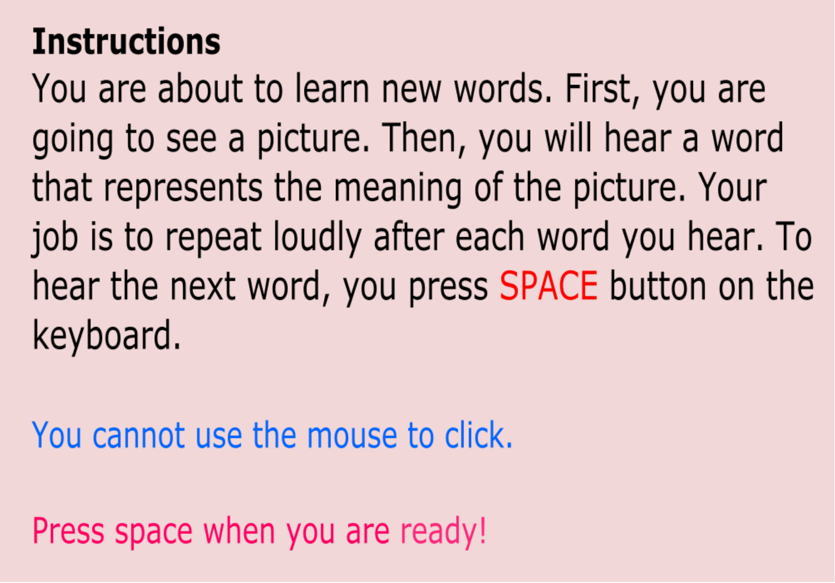
First, trainees were informed they would learn new vocabulary, and the meanings of the words would be shown in writing and represented by an image. They were not informed these were nonsense words, and the same wordlist was used in all three sessions. The training procedure took three consecutive days to complete. Pre-test recordings were not necessarily obtained right before the training sessions. However, post-test productions were always recorded directly after the third session of phonetic production training. The same procedures and devices used in recording training stimuli were followed to record trainees. The purpose of the pre-test was to obtain baseline productions that could be compared to the post-test productions to measure any improvement. In the pre-test, trainees were asked to read the words that appeared on the screen along with an image (see Figure 29). No further instructions were provided.

Figure 29. *An Example of What Was Represented on Screen to Trainees in the Pre- and Post-Tests*



After all trainees provided the pre-test productions, they started the training sessions online, but not necessary at the same day. The training sessions were done through the software PsyToolkit (Stoet, 2010, 2017) to minimize direct contact with trainees due to the pandemic. During each session, participants were monitored remotely via Zoom. No data were obtained from the training sessions. Since five English trainers provided the training sessions, five training experiments were designed. They all had the same images, but each experiment had the audio from one model talker. Thus, each trainer was provided with one experiment that had the audio from the model talker selected for the training sessions. First, the instructions appeared on the screen as shown in Figure 30.

Figure 30. *A Screenshot of Instructions Displayed on Psytoolkit for the Training Sessions*



Instructions
You are about to learn new words. First, you are going to see a picture. Then, you will hear a word that represents the meaning of the picture. Your job is to repeat loudly after each word you hear. To hear the next word, you press SPACE button on the keyboard.

You cannot use the mouse to click.

Press space when you are ready!

They were instructed to press the spacebar to move to the following trial after they finished saying the word they heard. For each trial, an image appeared on the screen while the corresponding word was heard through headphones 0.5 seconds later. Each word was repeated

12 times in each session. No instructions were given regarding how these words should be pronounced or what the intended segments were. They were asked to repeat the word they heard as clearly as possible but were not explicitly asked to imitate the trainer. They did two training sessions online while being monitored via Zoom by the researcher. The third session of training was done in person with the researcher.

The post-test recordings were obtained after the last training session. The same procedure used during the pre-test phase was followed. The same nonsense words with corresponding images were displayed on a screen. Their task was to read the word shown on a screen while being recorded. The aim of this phase was to collect the post-test productions that were analyzed to see how much their pronunciation of the target segments improved.

4.5 Pronunciation Improvement Assessment

The researcher anticipated that C-group would generally show more improvement in the trained segmental productions. It was also expected that if participants were able to converge phonetically to some native speakers, their pronunciation after being trained by the same native speakers would show greater intelligibility.

4.5.1 Listeners

A total of 40 native English listeners participated in the intelligibility judgement tasks (10 per task). Of these, 18 (45%) were male, and their ages ranged from 19 to 53 (*M* = 30.95, *SD* = 10.79). Listeners were recruited via Prolific.co, and each received a payment of $6.33. The raters from the perceptual similarity assessment in the phonetic convergence experiment were excluded. Listeners were prescreened to be monolingual English speakers born in the U.S., and those who failed attention check questions were excluded from the study. No hearing or speaking impairments were reported in the survey provided before they started the task.

4.5.2 Assessment Stimuli

The speech samples were the 12 nonsense words produced by the learners and model talkers. In addition, 12 filler (nontarget) words were added to minimize bias in categorizing segments. The resulting list consisted of 24 words in 12 minimal pairs (see Table 22 for the entire list). The words in each minimal pair varied by one segment expected to be confusable by Arabic speakers, such as the /p/ and /b/ in *puct* and *buct* or the /ɛ/ and /ɪ/ in *seft* and *sift*. Some nontarget words were real English words, such as *sift*, *mist*, and *suit*. Arabic speakers were asked to read the nontarget words after they provided the post-test productions so that their attention would not be drawn to the target segments. The 12 nontarget words produced by the trainers and trainees were not analyzed as they were used as fillers only.

Table 22. *Minimal Pairs Used in the Intelligibility Tasks*

Consonants				Vowels			
/p/ task		/v/ task		/ɛ/ task		/oʊ/ task	
/p/	/b/	/v/	/f/	/ɛ/	/ɪ/	/oʊ/	/u:/
puct	buct	vant	fant	seft	*sift*	roke	ruke
pusk	*busk*	vulb	fulb	mest	*mist*	sote	*suit*
paffs	baffs	vilt	filt	lesk	lisk	fote	fute

Note. Words in *italics* are real English words.

Trainees produced 480 target words in the pre- and post-test sessions (20 trainees x 12 target words x 2 sessions). Of the 120 words obtained from the five trainers, 60 formed the stimuli used in the training sessions (5 model talkers x 12 target words), while the other 60 represented the 12 nontarget words (5 model talkers x 12 nontarget words). In addition, 10 trainees were asked to produce the nontarget words, which resulted in 120 words. Not all trainees were asked to produce the nontarget words in order to make the intelligibility assessment more manageable for the English listeners. Four attention checks repeated twice were included in each task to ensure listeners were paying attention to the task and played all the sound files provided with each trial. The attention checks were straightforward. The labeling was the same within each task, but the audio file had a word obviously different from the other words. In the consonant intelligibility judgement tasks, listeners were presented with the words *teaser*, *cabbage*, *diesel*, and *dagger*. If any listener missed any one of the attention checks, their data were excluded from the experiment and they were not paid for their participation. In the vowel intelligibility judgment tasks, the attention checks included *sad*, *sleep*, *bad*, and *screen*, which had vowels that were obviously different from the target vowels.

In each intelligibility judgement task, there were 368 trials: (3 target words x 20 trainees x 2 times x 2 repetitions) + (3 target words x 5 trainers x 2 repetitions) + (3 nontarget words x 5 trainers x 2 repetitions) + (3 nontarget words x 10 trainees x 2 repetitions) + (4 attention checks x 2 repetitions). Thus, a total of 1,472 words were presented to the judges in the four intelligibility tasks. The native productions, trainee productions of nontarget words, and attention checks were not analyzed in the training results but were included in the reliability tests. Accordingly, for each task, 240 target nonsense words produced by the trainees during the pre- and post-tests were analyzed.

4.5.3 Assessment Procedure

To keep the assessment tasks to a manageable length for listeners, four judgment tasks were designed using PsyToolkit (Stoet, 2010, 2017). A summary of these tasks is provided in Table 23.

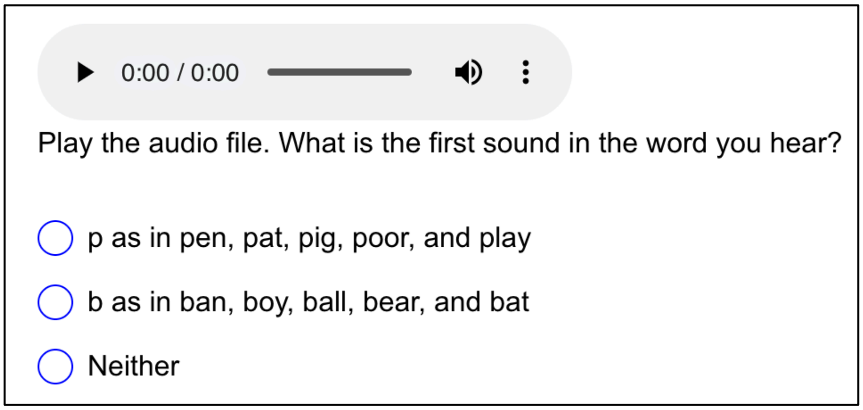
Table 23. *Summary of Participant Groups and Intelligibility Judgement Tasks*

Participant Group	Task
Speakers	
20 trainees	Pre- and post-production of 12 target and 12 nontarget words
5 trainers	Production of 12 target and 12 nontarget words
Listeners	
10 native English speakers	Intelligibility judgements for the target consonant /p/ (3 monosyllabic nonsense words) (/p/ task)
10 native English speakers	Intelligibility judgements for the target consonant /v/ (3 monosyllabic nonsense words) (/v/ task)
10 native English speakers	Intelligibility judgements for the target vowel /ɛ/ (3 monosyllabic nonsense words) (/ɛ/ task)
10 native English speakers	Intelligibility judgements for the target vowel /oʊ/ (3 monosyllabic nonsense words) (/oʊ/ task)

As mentioned previously, four experiments were constructed: two for target consonants and two for target vowels. Each of these experiments was presented to a different group of 10 native English speakers to evaluate the intelligibility of the target segments. For each task, 10 judges listened to the monosyllabic words recorded by the trainees during the pre- and post-tests. First, listeners were instructed to wear headphones while doing the task and were informed that the words were not necessarily real English words.

In the first task (/p/ task), listeners were presented auditorily with participants' productions of the words starting with /p/ and /b/. The labeling /p/ and /b/ were used because Arabic L2 learners of English tend to pronounce English /p/ as /b/. Listeners were asked to base their judgments only on the pronunciation of the first consonant. Three options appeared on the screen. The first was /p/, the second was /b/, and the third was *neither* (see Figure 31). Following Park and de Jong (2008), some examples of real English words containing these sounds were presented along with the labels to help listeners become familiar with the target sounds, particularly in the case of vowels. Thus, listeners' task was to categorize the sound they heard at the beginning of the word. They were able to repeat the stimuli as many times as they wanted. After they made their decision, a Next button activated so they could continue to the next trial, which was presented 600 milliseconds after they pressed the Next button. Within each task, trials were randomized. Listeners could take a break whenever they wanted as the task was self-paced. Each task took about 30–40 minutes to complete.

Figure 31. *A Screenshot of Instructions Displayed on Psytoolkit for the /p/ Task*



Play the audio file. What is the first sound in the word you hear?

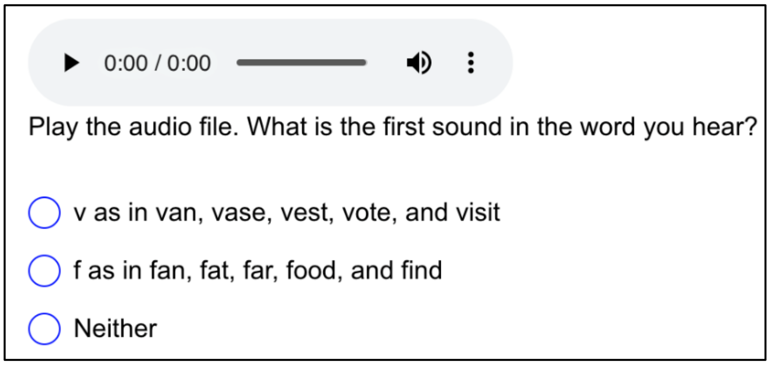
○ p as in pen, pat, pig, poor, and play

○ b as in ban, boy, ball, bear, and bat

○ Neither

The second intelligibility judgement task included the words beginning with /v/ and /f/. The procedure was the same as the /p/ task, except the labeling was different. The first option was /v/, the second was /f/, and the third was *neither* (see Figure 32). The labeling /v/ and /f/ was used as Arabic-speaking L2 learners of English tend to confuse English /v/ with /f/.

Figure 32. *A Screenshot of Instructions Displayed on Psytoolkit for the /v/ Task*



Play the audio file. What is the first sound in the word you hear?

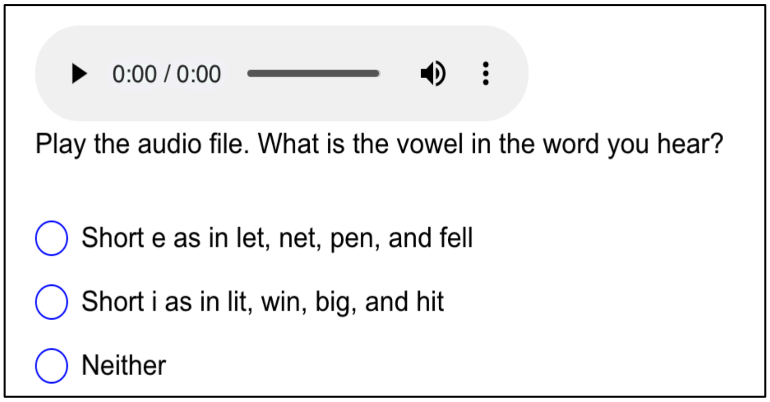
○ v as in van, vase, vest, vote, and visit

○ f as in fan, fat, far, food, and find

○ Neither

Following the same procedure, the third task contained the monosyllabic words with /ɛ/ and /ɪ/, while the fourth contained /oʊ/ and /u:/. Since listeners were not necessarily familiar with IPA symbols, symbols for English vowels that could be easily recognized by all listeners were used. *Short e* stood for /ɛ/ and *long o* stood for /oʊ/. Thus, the labels for the /ɛ/ task were *short e*, *short i*, and *neither* (see Figure 33), while those for the /oʊ/ task were *long o*, *long u*, and *neither*, as illustrated in Figure 34. Example words with the same vowels were provided along with each label. Again, the vowels /ɪ/ and /u/ were selected in the labeling as Arabic speakers tend to confuse them with /ɛ/ and /oʊ/, respectively.

Figure 33. *A Screenshot of Instructions Displayed on Psytoolkit for the /ɛ/ Task*



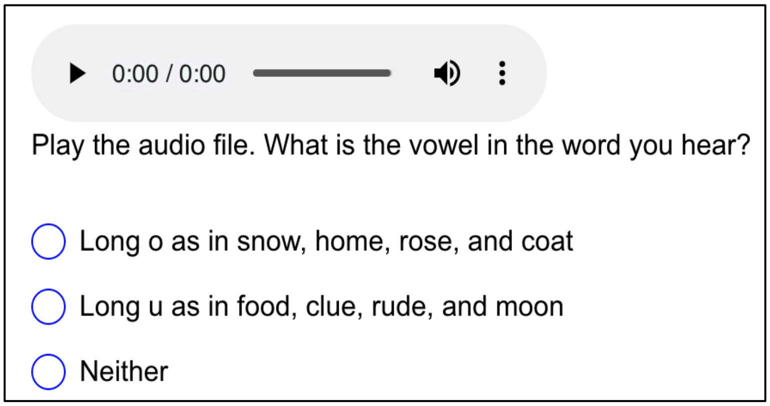
Play the audio file. What is the vowel in the word you hear?

○ Short e as in let, net, pen, and fell

○ Short i as in lit, win, big, and hit

○ Neither

Figure 34. *A Screenshot of Instructions Displayed on Psytoolkit for the /oʊ/ Task*



Play the audio file. What is the vowel in the word you hear?

○ Long o as in snow, home, rose, and coat

○ Long u as in food, clue, rude, and moon

○ Neither

4.5.4 Data Analysis

Fleiss' Kappa was used to test if the probability of raters' agreement was significantly above chance level, and an intraclass correlation test was used to ensure that the raters are consistent with one another. Then, three tests were carried out to examine the effects of training on intelligibility. Mixed between/within-subjects analysis of variance (a mixed-design ANOVA) was used to find differences between the groups' pre-test and post-test results. If any assumption of the ANOVA was violated, independent-samples *t*-tests were used instead. If any violation of this test's assumptions was found, a Mann–Whitney test was used as the alternative non-parametric test. Finally, a GEE model was run to examine the relationship between phonetic convergence on acoustic features and the magnitude of change from pre-test to post-test.

4.5.4.1 Reliability

As this study had four intelligibility judgement tasks (for the four target segments), four separate inter-rater reliability tests were carried out. All data obtained from the judges, including target and nontarget words, were used in the reliability tests. However, nontarget words were not examined in the subsequent analysis. First of all, Fleiss' Kappa showed the overall probability of agreement between judges for all tasks was above chance level ($p < .001$). After that, responses were counted as correct when they exactly matched the intended target segment. The proportion of correctly identified segments was calculated for each trainee in the pre- and post-test.

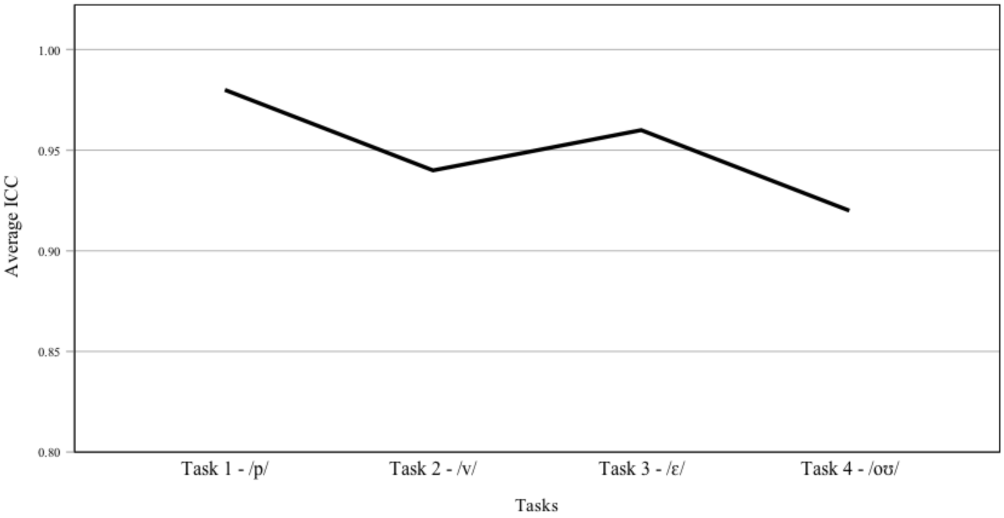
Inter-rater agreement was assessed for the 10 listeners for the /p/ task, and a high degree of reliability was found. The correlation of coefficients for ratings assigned to all words by the 10 raters in this task averaged .98 with a 95% confidence interval from .97 to .99 ($F(24,216) = 64.900$, $p < 0.\ 001$). For the /v/ task, a high degree of reliability was also found, although lower than the /p/ task, and the correlation of coefficients averaged .94 with a 95% confidence interval from .90 to .97 ($F(24,216) = 19.750$, $p < 0.001$). The correlation of coefficients for ratings assigned to the /ɛ/ task averaged .96 with a 95% confidence interval from .92 to .99 ($F(24,216) = 27.46$, $p < 0.001$). A high degree of reliability was found in the /oʊ/ task, but it scored slightly lower than the /ɛ/ task. The correlation of coefficients averaged .92 with a 95% confidence interval from .85 to .96 ($F(24,216) = 17.76$, $p < 0.001$). Table 24 and Figure 35 present the intraclass correlation coefficient (ICC) results. High reliability was found in all four intelligibility judgment tasks, although Task 4 had the lowest internal consistency (0.92).

Table 24. *Intraclass Correlation Coefficient for the Four Intelligibility Judgement Tasks*

		95% Confidence Interval					
	ICC	Lower Bound	Upper Bound	Value	*df1*	*df2*	*p*
Task 1 /p/	0.98	0.97	0.99	64.9	24	216	< 0.001
Task 2 /v/	0.94	0.90	0.97	19.75	24	216	< 0.001
Task 3 /ɛ/	0.96	0.92	0.98	27.46	24	216	< 0.001
Task 4 /oʊ/	0.92	0.85	0.96	17.76	24	216	< 0.001

Note. 95% Lower/upper CI = 95% confidence interval of ICC.

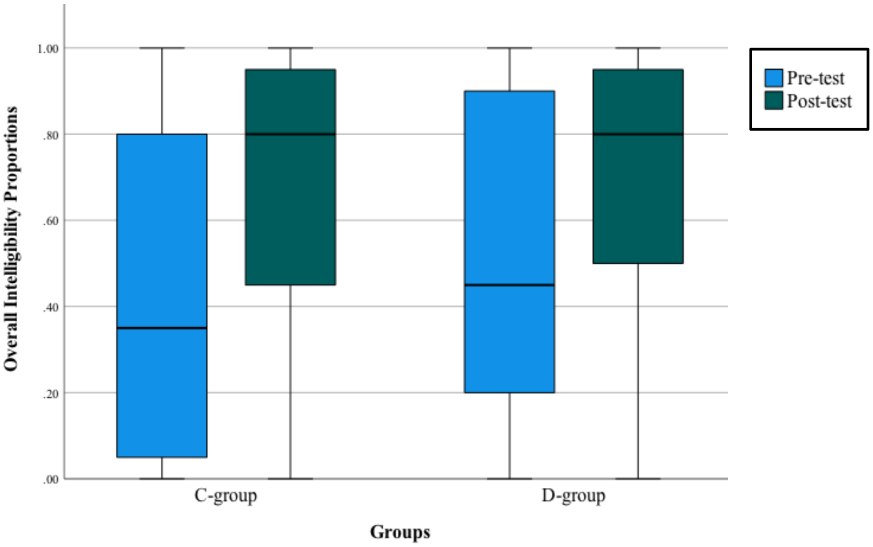
Figure 35. *Average Intraclass Correlation Coefficient by Intelligibility Judgement Task*



4.5.4.2 Intelligibility Improvement across the Four Target Segments

This section reports the changes in trainees' overall performance from pre- to post-test. It begins with the overall intelligibility improvement collapsed across the four training segments, before discussing the results per task in more detail. Segment intelligibility improvement was first calculated as the number of stimuli correctly identified by judges as the target segment. Then, intelligibility scores were computed as the proportion of correct identification in the pre- and post-tests averaged across the four target segments (i.e., /p, v, ɛ, oʊ/). Since the aim of this section is to examine if the training method used in the study was effective, only data obtained from the training segments in the two groups have been analyzed. Therefore, the intelligibility scores for the native productions and the nontarget words were excluded. Figure 36 illustrates the proportions of correct identification across all segments from the C-group and D-group as assessed by the judges.

Figure 36. *Intelligibility Proportions Averaged across the Target Segments for C-Group and D-Group*



As can be seen in Figure 36, both groups' productions showed substantial improvement in intelligibility from pre-test to post-test. To examine the training's effect on overall intelligibility, mean proportions of the groups' pre- and post-tests were compared using a mixed-design ANOVA. The between-group factor was group type (C-group and D-group), while time (pre- and post-test) was the within-group factor. No violation of the assumptions of homoscedasticity (i.e., assumption of equal variances) or homogeneity of variances was found, $p > .05$. A Kolmogorov-Smirnov test also revealed that the data were normally distributed. The results showed no significant interaction between the two groups and time, Wilks' Lambda = .96, $F(1,18) = .854$, $p > .05$, partial eta squared = .045. There was, however, a significant main effect for time (Wilks' Lambda = .62, $F(1,18) = 10.90$, $p = 004$, partial eta squared = .377), with both groups showing an increase in intelligibility scores across the two time periods (see Table 25). The main effect comparing the two groups was not significant ($F(1,18) = 1.050$, $p > .05$, partial eta squared = .055), suggesting no difference between the two groups' intelligibility scores from the pre- and post-tests.

Table 25. *Overall Intelligibility Proportions for C-Group and D-Group in the Pre-Test and Post-Test*

Test	Group	*M*	*SD*	*N*
Pre-test	C-group	.41	.18	10
	D-group	.52	.10	10
Post-test	C-group	.68	.13	10
	D-group	.69	.12	10

Although the analysis did not show a significant difference between the groups' intelligibility improvement after training, the magnitude of change was higher for C-group (*M* = .27, *SD* = .16), as illustrated in Figure 37. Means, medians, and standard deviations are displayed in Table 26. The Mann–Whitney test revealed no significant difference in the magnitude of change for C-group (*Md* = .20, *SD* =.16) and D-group (*Md* = .10, *SD* =.13), *U* = 6458.50, *z* = -1.38, *p* = .17, *r* = -0.31. These results suggest the groups performed similarly overall.

Figure 37. *Magnitude of Overall Change from Pre-Test to Post-Test in the Two Groups*

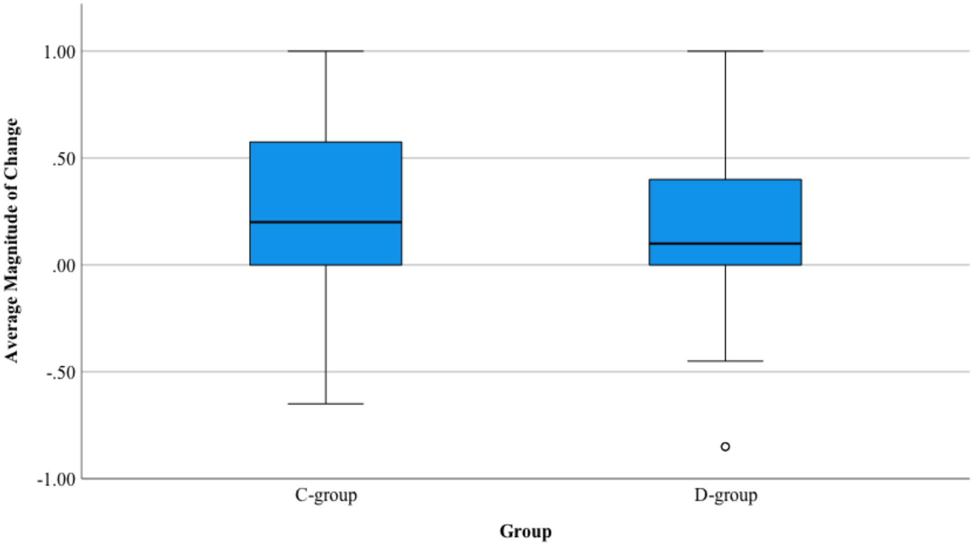
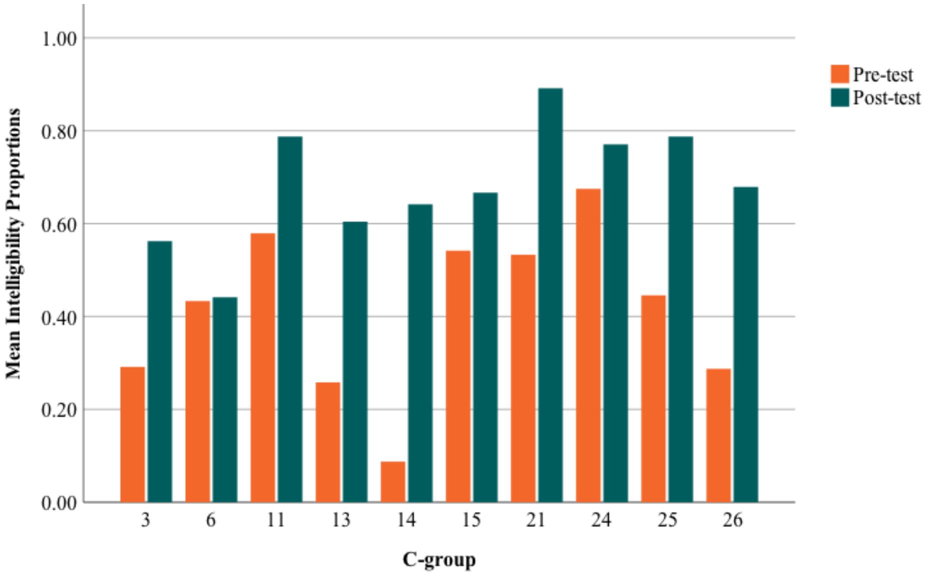


Table 26. *Descriptive Statistics of Overall Change from Pre-Test to Post-Test in the Two Groups*

Group	*N*	*M*	*Md*	*SD*
C-group	10	.27	.20	.16
D-group	10	.18	.10	.13

Figure 38 shows the magnitude of change in intelligibility scores for each of the 10 trainees in C-group according to the judges. All trainees in C-group, except for Trainees 6 and 24, showed considerable improvement from pre- to post-test; however, there were discrepancies among trainees and in the magnitude of change. The magnitude of change was calculated as the post-test minus the pre-test. Positive differences indicated that the intelligibility scores improved from pre- to post-test, and negative values indicated the trainee's performance decreased in the post-test. For example, Trainee 14 exhibited the highest increase by over 50% in her intelligibility score from pre-test to post-test, while Trainees 13, 21, 25, and 26 improved by over 30%. Trainees 3 and 11 increased by over 20%, and Trainee 3 only increased by about 10%. The only one to not show any intelligibility improvement was Trainee 6. See Appendix C for individual trainees' overall intelligibility scores.

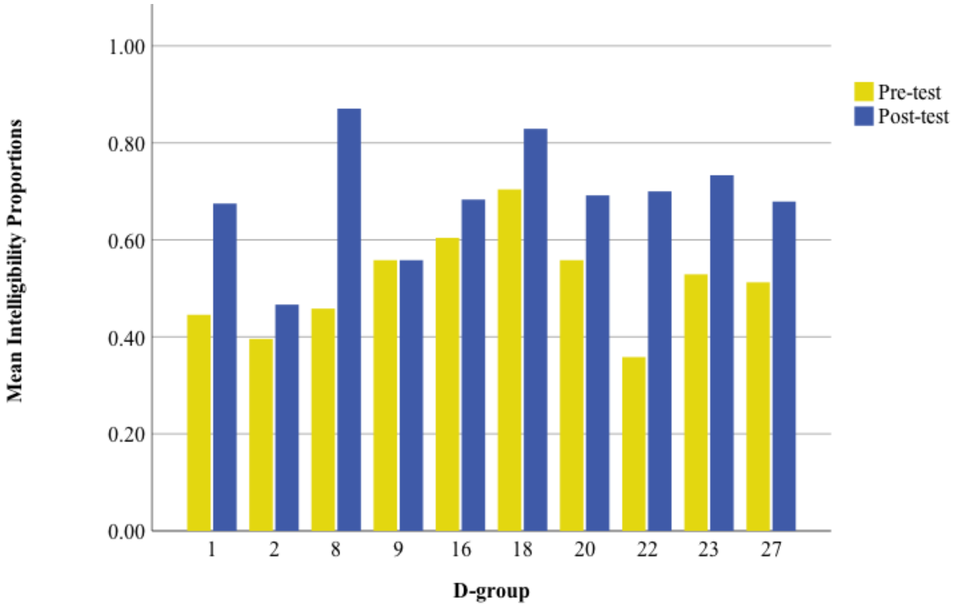
Figure 38. *C-Group Intelligibility Improvement Scores Averaged across the Four Target Segments*



Note. The x-axis displays each trainee's intelligibility scores on the pre- and post-tests. The y–axis represents the scores as a proportion of correct identification of all segments.

Figure 39 illustrates the magnitude of change in intelligibility scores for each of the 10 trainees in D-group according to the judges. For the individual trainees' intelligibility scores in D-group, see Appendix D. The majority of trainees showed more intelligible segments after training, but generally less so than C-group. Trainee 8 showed the most improvement by 41%, which was still smaller than the highest degree of improvement achieved by Trainee 14 (50%) in C-group. Trainee 22 exhibited around 30% improvement, Trainees 2 and 16 showed less than 10%, and Trainee 9 did not show any improvement after training.

Figure 39. *D-Group Intelligibility Improvement Scores Averaged across the Four Target Segments*



Note. The x-axis displays each trainee's intelligibility scores on the pre- and post-tests. The y–axis represents the scores as a proportion of correct identification of all segments.

4.5.4.3 Intelligibility Improvement in the Production of /p/

To examine the effect of training on the intelligibility of the target segment /p/, intelligibility scores were computed as the proportion of correct identification. Thus, for each trainee, the number of responses correctly identified by listeners as the target segment were counted for the pre- and post-tests. Then, the proportion of correct responses was calculated. The same data analysis was followed for each segment separately. Figure 40 displays proportions of correctly identified /p/ by C-group and D-group. Both groups were perceived as more intelligible on the post-test. This observation was tested using a mixed-design ANOVA, which revealed no significant interaction between the two groups and time, Wilks' Lambda = 1.000, *F*(1,18) = 000, *p* > .05, partial eta squared = 000. However, a significant main effect for time was found, Wilks'

Lambda = .52, *F*(1,18) = 12.231, *p* = .003, partial eta squared = 405, with both groups showing an increase in intelligibility scores across the two time periods (see Table 27). The main effect comparing the two groups was not significant (*F*(1,18) = 7.0, *p* > .05, partial eta squared = .04), indicating the groups were not significantly different in /p/ intelligibility scores in the pre- and post-tests.

Figure 40. *C-Group and D-Group Intelligibility Proportions for /p/*

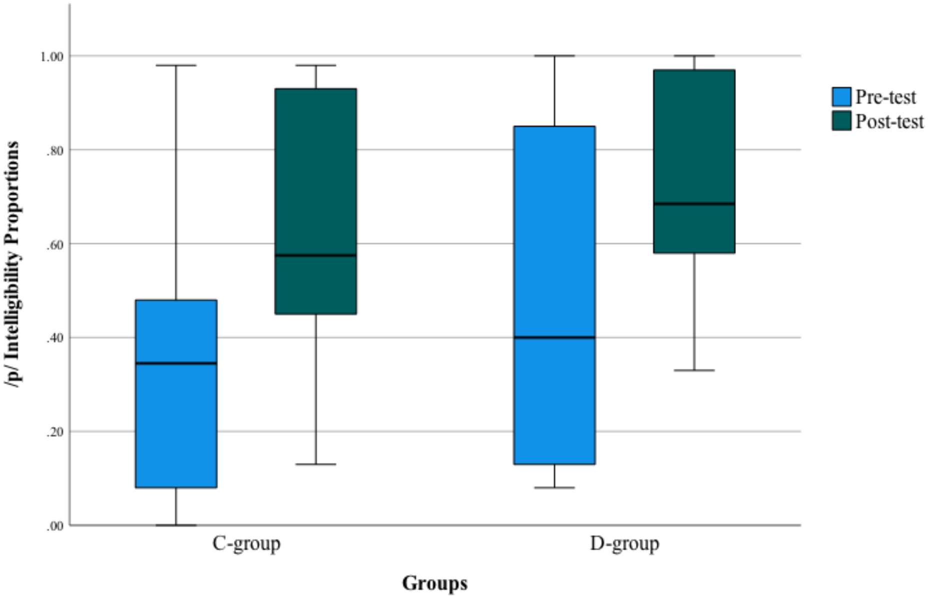


Table 27. *Descriptive Statistics of /p/ Intelligibility Proportions for C-Group and D-Group*

Test	Group	*M*	*SD*	*N*
Pre-test	C-group	.39	.34	10
	D-group	.49	.34	10
Post-test	C-group	.62	.28	10
	D-group	.72	.22	10

Figure 41 illustrates the magnitude of change from pre- to post-test in terms of /p/ intelligibility for both groups. Statistical analysis did not show any significant difference between the groups; however, D-group's box plot is taller than C-group's box plot, suggesting D-group showed more variation in intelligibility improvement for /p/. In addition, some D-group trainees exhibited higher improvement than trainees in C-group, which was unexpected. Descriptive statistics are displayed in Table 28.

Figure 41. *Magnitude of Change from Pre- to Post-Test /p/ Intelligibility Scores in the Two Groups*

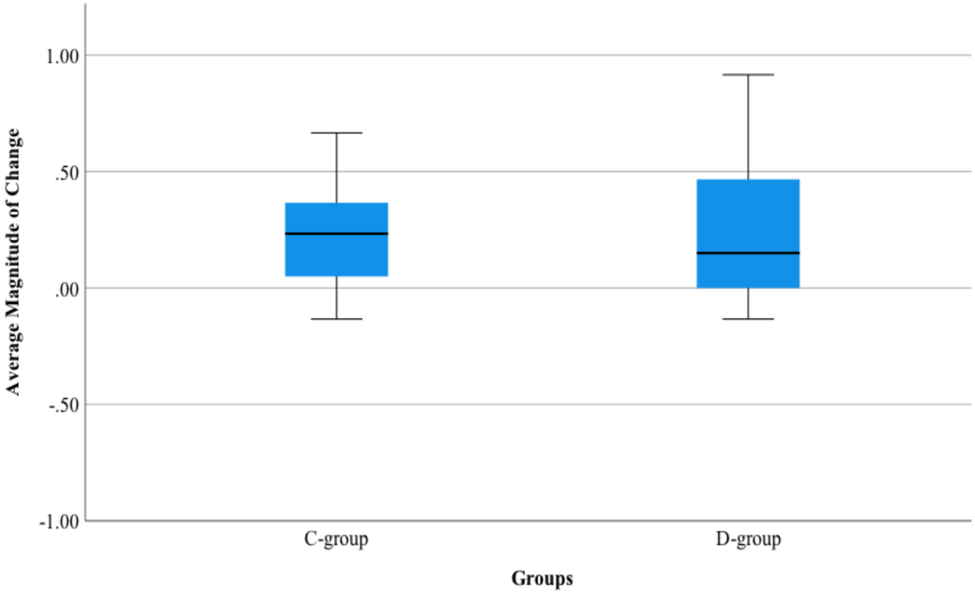
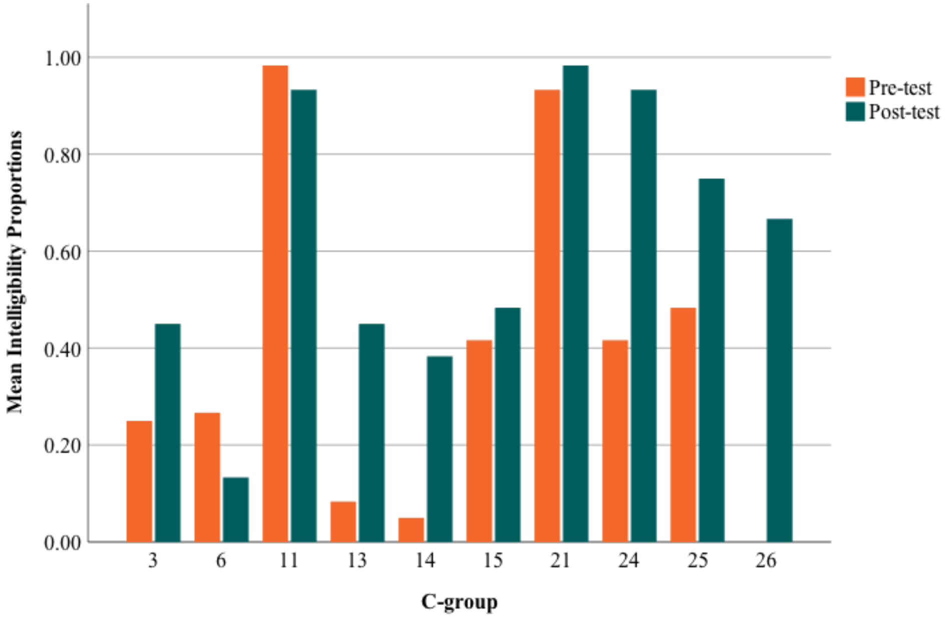


Table 28. *Intelligibility Score Change from Pre- to Post-Test for /p/*

Group	*N*	*M*	*Md*	*SD*
C-group	10	.23	.23	.25
D-group	10	.23	.15	.32

There did not appear to be much difference between groups' intelligibility performance after the training, and some trainees' performance decreased slightly. Figure 42 shows the magnitude of change found in intelligibility scores for each of the 10 trainees in C-group. Except for Trainees 6 and 11, C-group was perceived as more intelligible on the post-test, but there were differences across trainees. Their pre-test scores ranged widely from 0% to 98%. Some (i.e., Trainees 11 and 21) produced a highly intelligible /p/ before training, leaving little room for improvement. Trainees 3, 13, 14, 24, 25, and 26 substantially increased from pre- to post-test, with Trainee 24 exhibiting the highest performance. Trainee 15 only increased a little, while Trainee 6 showed the poorest performance. See Appendix E for more details about the individual intelligibility scores for /p/ in C-group.

Figure 42. *C-Group Intelligibility Score Improvement for /p/*

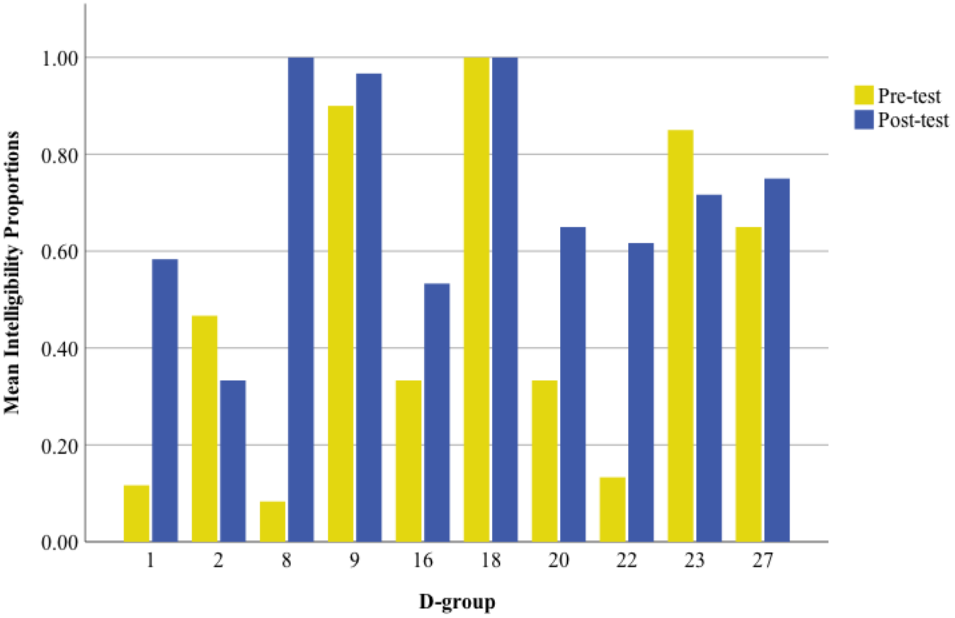


Note. The x-axis displays each trainee's pre- and post-test intelligibility scores. The y-axis represents the scores as a proportion of correct identification of /p/.

D-group's /p/ intelligibility scores are illustrated in Figure 43. Trainees 1, 8, 9, 16, 20, and 22 produced a more intelligible /p/ after training, with Trainee 8 showing the highest improvement. Although the majority of trainees were perceived as having a more intelligible /p/

on the post-test, there were considerable differences as pre-test scores ranged from 8% to 100%. Some (i.e., Trainees 9, 18, 23, and 27) produced a highly intelligible /p/ before training, leaving little room for improvement. However, the performance of some dropped after training (see Appendix F for more details).

Figure 43. *D-Group Intelligibility Score Improvement for /p/*



Note. The x-axis displays each trainee's pre- and post-test intelligibility scores. The y-axis represents the scores as a proportion of correct identification of /p/.

4.5.4.4 Intelligibility Improvement in the Production of /v/

Means, medians, and standard deviations for /v/ are displayed in Table 29. Figure 44 shows the proportions of correct identification of /v/ in the two groups according to the judges.

To assess the impact of training on the intelligibility of /v/, a mixed-design ANOVA could not be used due to violations of the assumptions of homoscedasticity and homogeneity of variances, $p < .05$. The data were also not normally distributed. Descriptive statistics for changes from pre- to post-test are shown in Table 30. A Mann–Whitney test on the magnitude of change between pre- and post-test scores found no significant difference between C-group (*Md* = .00, *SD* = .26) and D-group (*Md* = -.02, *SD* = .32), $U = 32.00$, $z = -1.36$, $p = .17$, $r = -0.30$. The magnitude of change in the two groups is illustrated in Figure 45.

Table 29. *Descriptive Statistics of /v/ Intelligibility Proportions for C-Group and D-Group*

Test	Group	*M*	*MD*	*SD*	*N*
Pre-test	C-group	.74	.88	.34	10
	D-group	.90	.93	.10	10
Post-test	C-group	.89	.91	.10	10
	D-group	.84	.92	.17	10

Figure 44. *C-Group and D-Group Intelligibility Proportions for /v/*

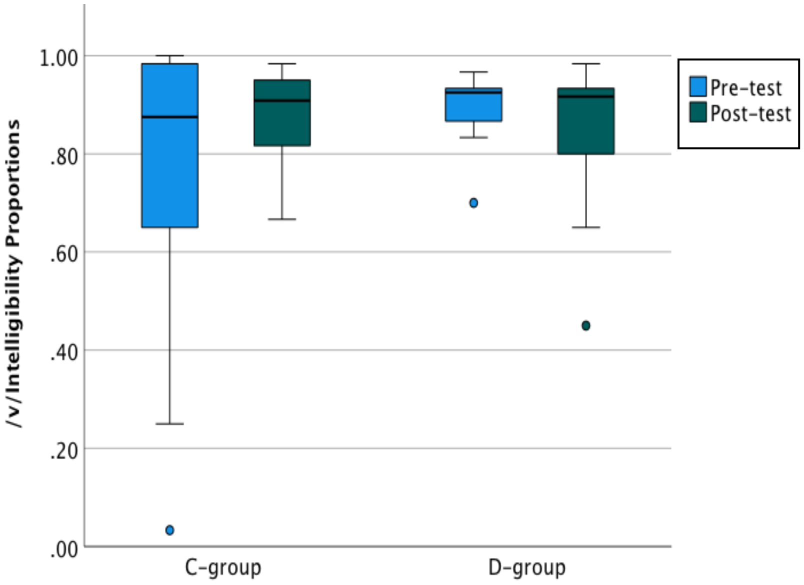
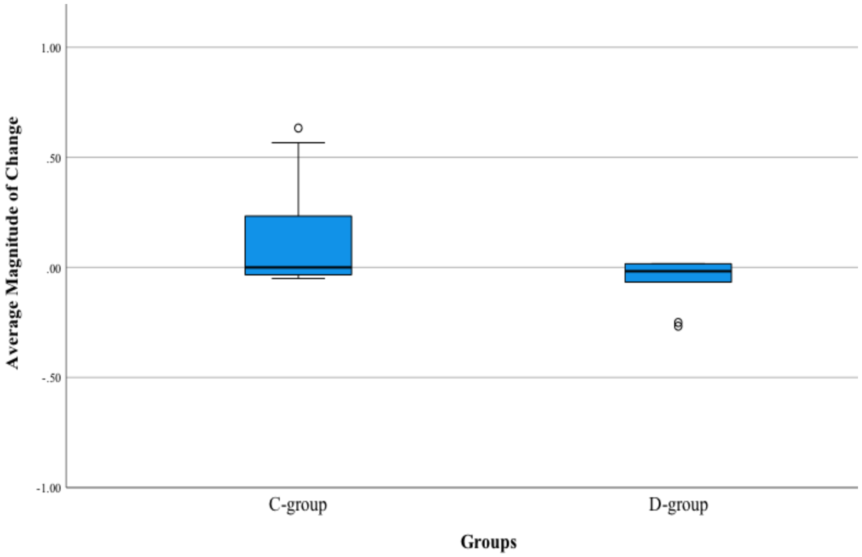


Table 30. *Intelligibility Score Change from Pre- to Post-Test for /v/*

Group	*N*	*M*	*Md*	*SD*
C-group	10	.15	.00	.26
D-group	10	-.10	-.02	.32

Another two Mann–Whitney tests were conducted on the pre- and post-test scores to examine if the two groups' performance was significantly different from one another at any time. However, no significant difference was found between C-group (*Md* = .88, *SD* = .34) and D-group (*Md* = .93, *SD* = 10) in the pre-test, *U* = 57.5, *z* = .57, *p* = .57, *r* = .13. There was also no significant difference between C-group (*Md* = .91, *SD* = .10) and D-group (*Md* = .92, *SD* = 17) in the post-test, *U* = 39.5, *z* = -.80, *p* > .05, *r* = -.18.

Figure 45. *Magnitude of Change from Pre- to Post-Test /v/ Intelligibility Scores in the Two Groups*

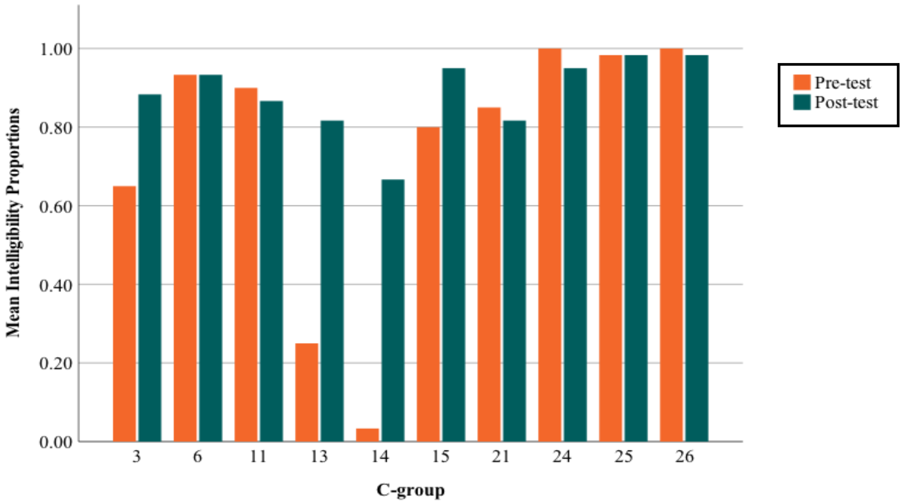


A related-samples Wilcoxon signed rank test revealed no significant differences in the median of intelligibility scores for C-group between the pre-test (*Md* = .88, *SD* = .34) and post-test (*Md* = .91, *SD* = .10), *Z* = 26.00, *p* > .05. The median of intelligibility scores for D-group on the pre-test (*Md* = .93, *SD* = .10) was likewise not significantly different from the post-test (*Md* = .84, *SD* = .17), *Z* = 10.00, *p* > .05.

Although the statistical analysis showed no significant differences between groups at any level, the spread of data was different. As Figure 44 shows, there was a greater variability in mean /v/ intelligibility scores in C-group on the pre-test than in D-group as well as a larger outlier in the two groups. The range of data in D-group was comparatively small, suggesting trainees' overall performance was relatively comparable. On the post-test, the range of data was relatively similar between the two groups. C-group's post-test box plot was shorter and higher than the one for the pre-test, indicating most trainees produced a more intelligible /v/ after training. D-group's post-test box plot was lower than the pre-test, indicating that some scores dropped after training. That could be attributed to their high performance on the pre-test leaving little room for improvement.

Figures 46 and 47 show individual /v/ intelligibility scores for C-group and D-group, respectively. Most trainees in C-group produced a highly intelligible /v/ on the pre-test. Only Trainee 13 (25%) and Trainee 14 (3%) showed low performance on the pre-test. Therefore, the training was shown to be highly effective in this regard.

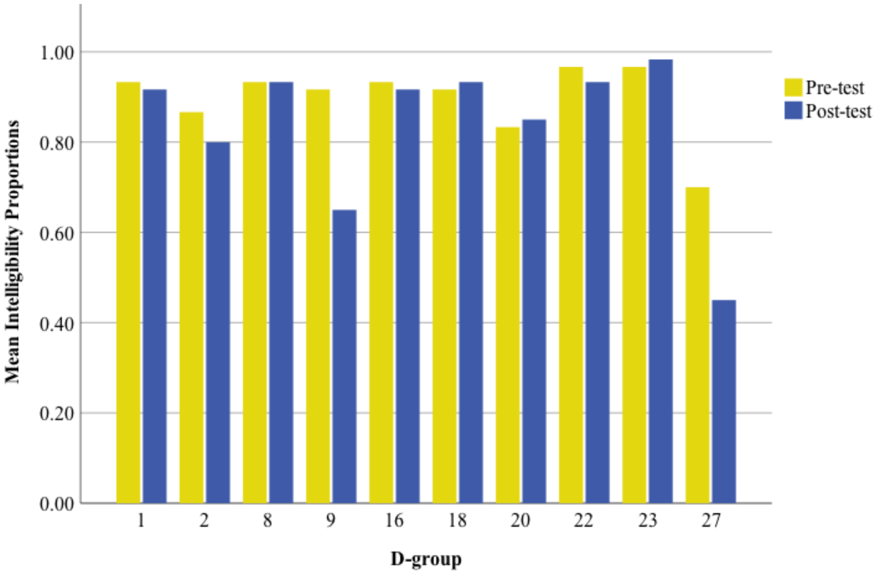
Figure 46. *C-Group Intelligibility Score Improvement for /v/*



Note. The x-axis displays each trainee's pre- and post-test intelligibility scores. The y-axis represents the scores as a proportion of correct identification of /v/.

The performance of some trainees in D-group (i.e., Trainees 24 and 26) went down slightly as they always produced /v/ correctly on the pre-test. D-group, except Trainees 20 and 27, produced a highly intelligible /v/ at a level higher than 90%. Productions by Trainee 20 (83%) and Trainee 27 (70%) were also identified as highly intelligible, but to a lower level. More than half of D-group showed a slight decrease on the post-test, as there was little room for improvement. For the complete list of individual /v/ intelligibility scores for both groups, see Appendices G and H.

Figure 47. *D-Group Intelligibility Score Improvement for /v/*



Note. The x-axis displays each trainee's pre- and post-test intelligibility scores. The y-axis represents the scores as a proportion of correct identification of /v/.

4.5.4.5 Intelligibility Improvement in the Production of /ɛ/

Figure 48 shows the proportions for correct identification of /ɛ/ by the judges. A mixed-design ANOVA, with group type (C-group and D-group) as the between-subject factor and time (pre- and post-test) as the within-subject factor, showed no significant interaction between the two groups and time, Wilks' Lambda = .925, *F*(1,18) = 1.46, *p* > .05, partial eta squared = .08. There was, however, a substantial main effect for time (Wilks' Lambda = .39, *F*(1,18) = 27.87, *p* < .001, partial eta squared = .61), with both groups doing significantly better in the post-tests (see Table 31). Such results suggested that shadowing indeed helped trainees produce a more intelligible /ɛ/, regardless of group. The main effect comparing the two groups was not significant (*F*(1,18) = 2.29, *p* > .05, partial eta squared = .11), suggesting the two groups did not differ in terms of their intelligibility scores across the pre- and post-tests.

Figure 48. *C-Group and D-Group Intelligibility Proportions for /ɛ/*

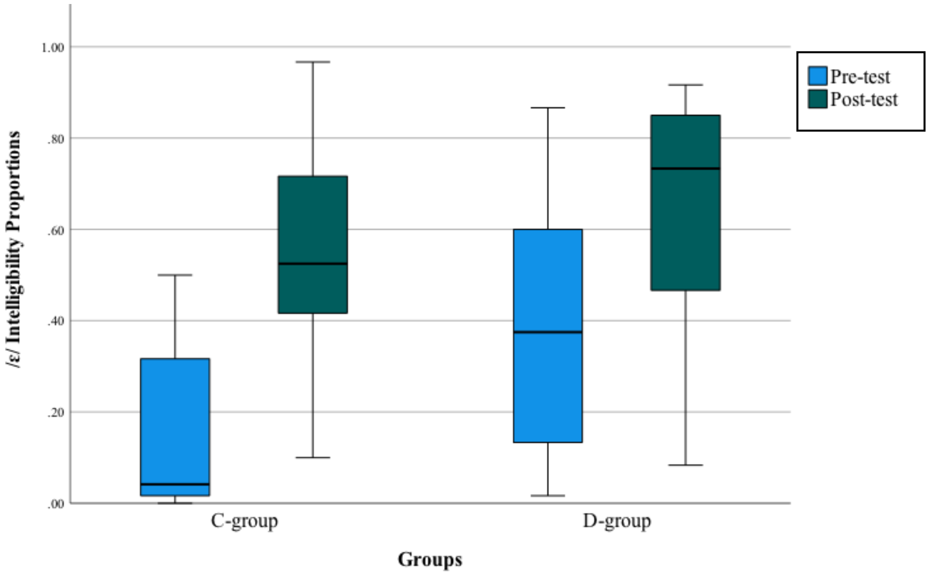


Table 31. *Descriptive Statistics of /ɛ/ Intelligibility Proportions for C-Group and D-Group*

Test	Group	*M*	*SD*	*N*
Pre-test	C-group	.15	.19	10
	D-group	.37	.30	10
Post-test	C-group	.54	.25	10
	D-group	.62	.30	10

The statistical analysis showed no significant differences between groups at any level, but the spread of data was different. D-group showed a greater variability in pre-test /ɛ/ intelligibility scores than C-group. The range of data in C-group was relatively small, suggesting trainees' performance was relatively comparable. In D-group, the distribution of data was more spread out in the pre-test, and the box plot was higher than C-group's box plot. This indicated that D-group performed better than C-group on the pre-test.

On the post-test, the groups showed a substantial increase in intelligibility, and their range of data was similar. The groups' box plots were higher than the ones for the pre-test, indicating most trainees produced a more intelligible /ɛ/ after training. However, as illustrated in Figure 49, the magnitude of change was higher for C-group (*M* = .39, *SD* = .32). Means, medians, and standard deviations of change from pre- to post-test for the /ɛ/ task are displayed in Table 32.

Figure 49. *Magnitude of Change from Pre- to Post-Test /ɛ/ Intelligibility Scores in the Two Groups*

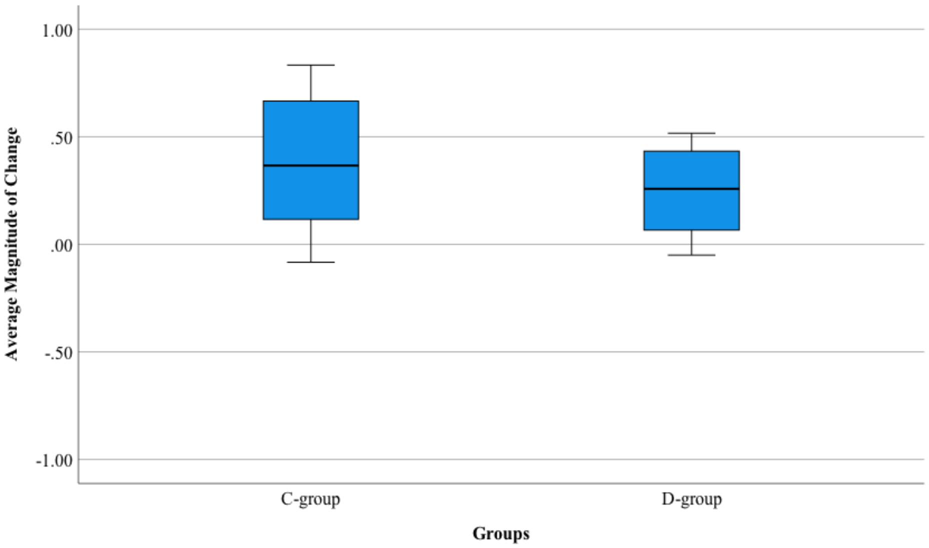
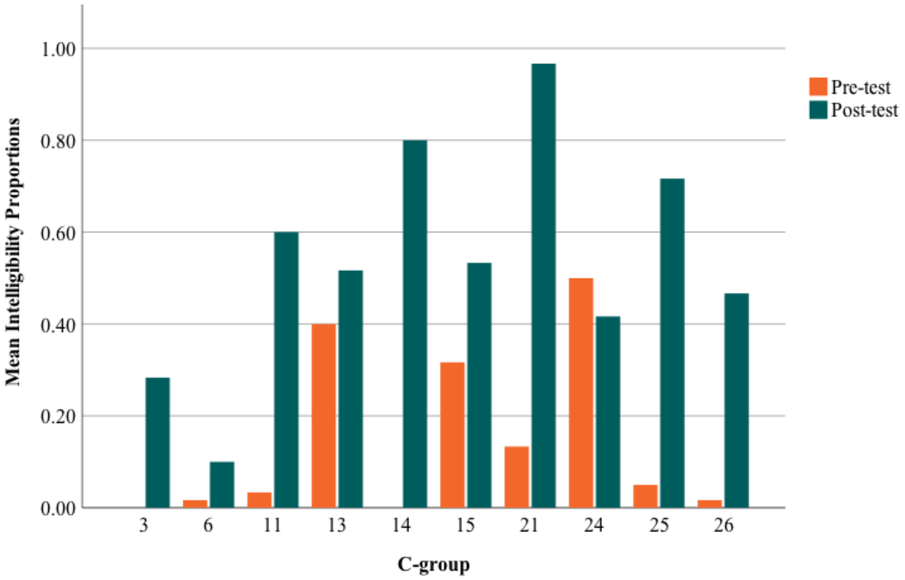


Table 32. *Intelligibility Score Change from Pre- to Post-Test for /ɛ/*

Group	*N*	*M*	*Md*	*SD*
C-group	10	.39	.37	.32
D-group	10	.25	.26	-.05

Figures 50 and 51 show individual /ɛ/ intelligibility scores for C-group and D-group in the pre- and post-tests. In C-group, 70% produced a poorly intelligible /ɛ/ on the pre-test, and less than 10% of their productions were identified correctly. The highest performance was scored by Trainee 24 (50%). Training sessions were effective in improving the intelligibility of most trainees in C-group. Only Trainee 24 showed a slight decrease after training, whereas Trainees 6 and 13 showed a small improvement.

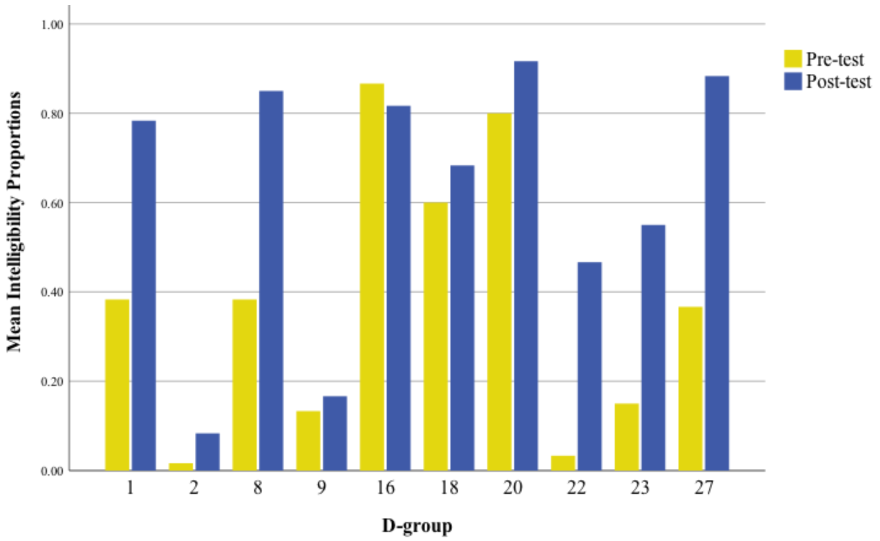
Figure 50. *C-Group Intelligibility Score Improvement for /ɛ/*



Note. The x-axis displays each trainee's pre- and post-test intelligibility scores. The y-axis represents the scores as a proportion of correct identification of /ɛ/.

Although the statistical analysis did not show any significant differences between groups, D-group produced a more intelligible /ɛ/ on the pre-test than C-group. Trainees 16, 18, and 20 produced a highly intelligible /ɛ/ and thus could not show much improvement. Trainee 2 had a very poor performance on the pre-test and showed no improvement after training. For individual trainees' pre-test and post-test intelligibility scores in the /ɛ/ task, see Appendices I and J.

Figure 51. *D-Group Intelligibility Score Improvement for /ɛ/*



Note. The x-axis displays each trainee’s pre- and post-test intelligibility scores. The y-axis represents the scores as a proportion of correct identification of /ɛ/.

4.5.4.6 Intelligibility Improvement in the Production of /oʊ/

Figure 52 shows the proportions for correct identification of /oʊ/ by the judges. To examine the training’s effect on /oʊ/ intelligibility, a mixed-design ANOVA could not be used due to violations of the assumptions of homoscedasticity and homogeneity of variances, $p < .05$. Descriptive statistics for changes from pre- to post-test are shown in Table 33. To examine if the two groups’ performance was significantly different, an independent-samples *t*-test was conducted on the pre- and post-test scores, as the data were normally distributed. No significant difference was found between C-group ($M = .38$, $SD = .21$) and D-group ($M = .29$, $SD = .07$) on the pre-test, $t(18) = 1.23$, $p > .05$. However, there was a significant difference between C-group ($M = .69$, $SD = .07$) and D-group ($M = .58$, $SD = .15$) on the post-test, $t(18) = 2.07$, $p = .05$, two-tailed. This meant C-group did significantly better than D-group after training.

Figure 52. *C-Group and D-Group Intelligibility Proportions for /oʊ/*

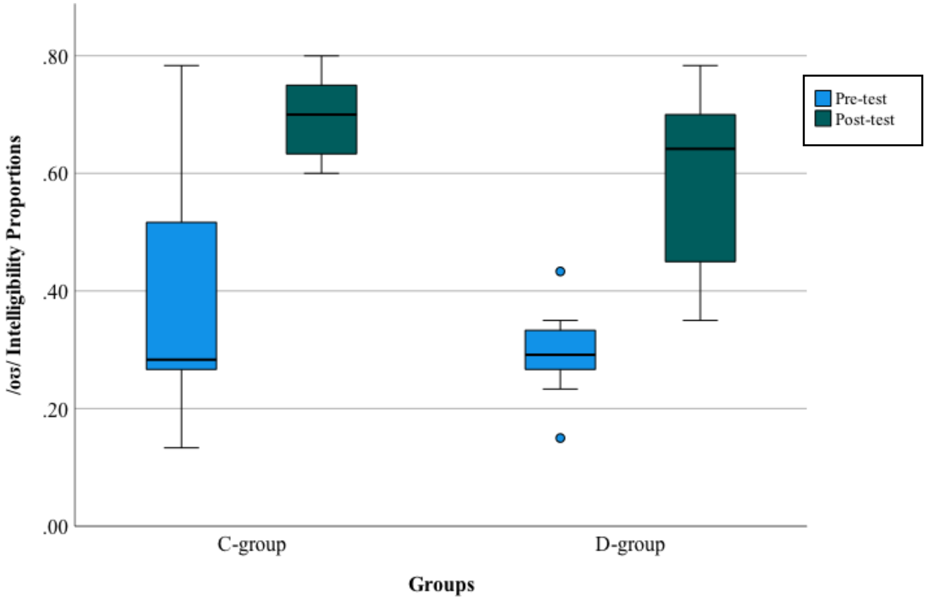


Table 33. *C-Group and D-Group Intelligibility Proportions for /oʊ/*

Test	Group	*N*	*M*	*SD*
Pre-test	C-group	10	.38	.21
	D-group	10	.29	.07
Post-test	C-group	10	.69	.07
	D-group	10	.58	.15

A paired-sample *t*-test was conducted to evaluate the effectiveness of training within each group. There was a significant increase in intelligibility from the pre-test (*M* = .14, *SD* = .18) to the post-test (*M* = .54, *SD* = .25) in C-group, *t*(9) = 3.91, *p* = .004, two-tailed. The mean increase in intelligibility scores was .39 with a 95% confidence interval ranging from .16 to .62. The eta squared (.32) indicated a small effect size. A paired-sample *t*-test for D-group likewise found a significant increase in intelligibility scores from the pre-test (*M* = .37, *SD* = .20) to the post-test (*M* = .62, *SD* = .29), *t*(9) = 3.64, *p* = .005, two-tailed. The mean increase in intelligibility scores was .24 with a 95% confidence interval ranging from .09 to .39. The eta squared (.21) indicated a small effect size. These results suggested that training was effective in improving the intelligibility of /oʊ/ in both groups.

Another independent-samples *t*-test was conducted on the differences (i.e., magnitude of change) between the pre- and post-test scores between groups, as the data were normally distributed. No significant difference was found in the magnitude of change for C-group (*M* = .31, *SD* = .20) and D-group (*M* = .29, *SD* = .16), *t*(18) = .288, *p* > .05, two-tailed. The magnitude of change in the two groups is illustrated in Figure 53, with means and standard deviations given in Table 34. The range of change in C-group was more spread out than D-group, indicating the /oʊ/ intelligibility improvement in C-group had more variation, though the average change was comparable.

Figure 53. *Magnitude of Change from Pre- to Post-Test /oʊ/ Intelligibility Scores in the Two Groups*

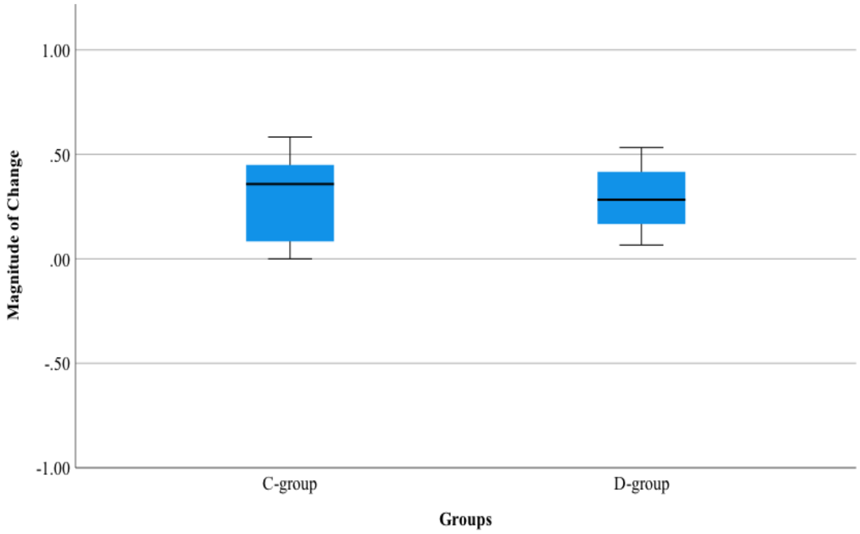
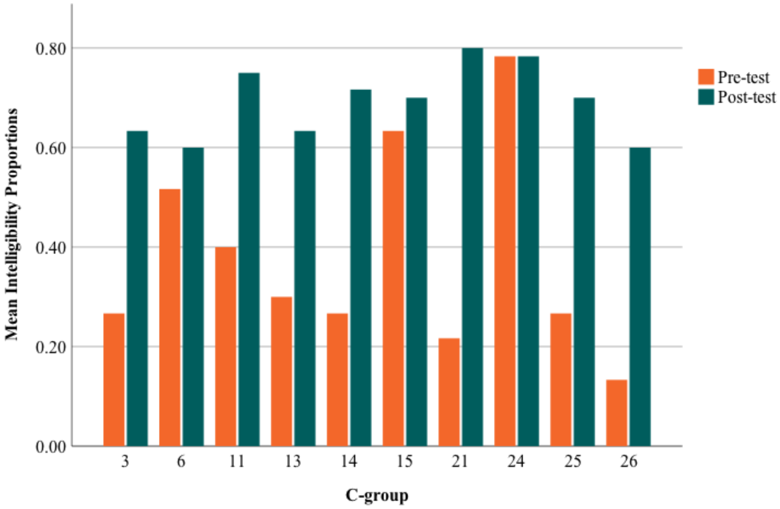


Table 34. *Intelligibility Score Change from Pre- to Post-Test for /oʊ/*

Group	*N*	*M*	*SD*
C-group	10	.31	.20
D-group	10	.29	.16

Figures 54 and 55 show the pre-test and post-test individual /oʊ/ intelligibility scores for C-group and D-group. In C-group, considerable differences were observed across their intelligibility scores on the pre-test. Trainees 15 and 24 exhibited a highly intelligible /oʊ/ in the pre-test, leaving little room for improvement. Others had productions identified correctly around 30% of the time (i.e., Trainees 3, 13, 14, and 25). After training sessions, 70% of trainees showed considerable intelligibility improvement. The highest magnitude of improvement was scored by Trainee 21. Hence, training sessions were effective in improving the intelligibility performance of most trainees in C-group.

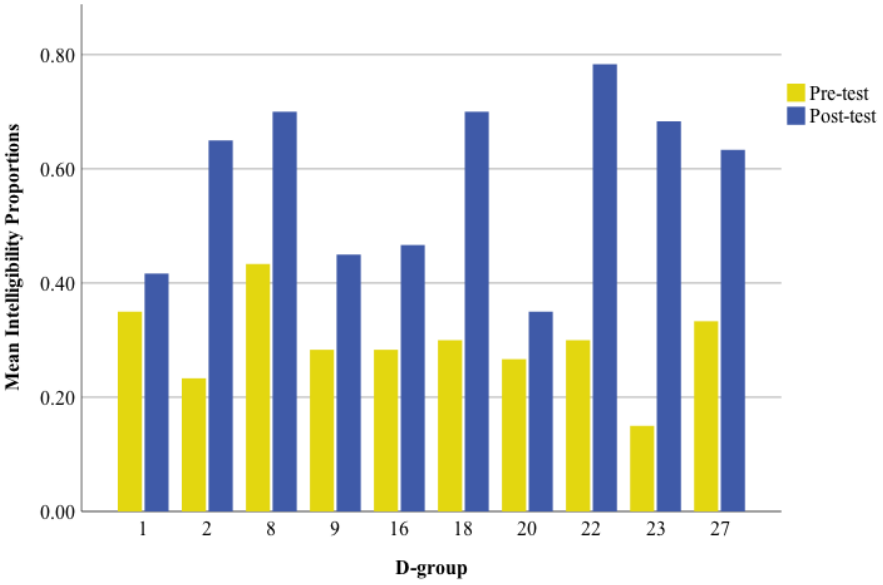
Figure 54. *C-Group Intelligibility Score Improvement for /oʊ/*



Note. The x-axis displays each trainee’s pre- and post-test intelligibility scores. The y-axis represents the scores as a proportion of correct identification of /oʊ/.

D-group also showed considerable variation across their intelligibility scores on the pre-test. Trainee 8 produced the most intelligible /oʊ/ on the pre-test, but it was lower than the highest pre-test scores in C-group. Some (i.e., Trainees 2, 18, 22, and 23) showed a drastic increase, while others increased to a lesser degree (i.e., Trainees 8 and 27). The intelligibility scores of some trainees increased very slightly by less than 10% (i.e., Trainees 1 and 20). For more details about individual trainees' intelligibility scores in the /oʊ/ task, see Appendices K and L.

Figure 55. *D-Group Intelligibility Score Improvement for /oʊ/*



Note. The x-axis displays each trainee's pre- and post-test intelligibility scores. The y-axis represents the scores as a proportion of correct identification of /oʊ/.

4.5.4.7 Overall

To determine if any group showed more improvement than the other group in a particular segment, a Kruskal-Wallis test compared differences in the magnitude of change in intelligibility scores across the four target segments within each group. For C-group (χ^2 (3, 40) = 5.13, p >.05), no significant difference was found in the degree of improvement among the four segments. For D-group, there was a significant difference in two target segments (χ^2 (3, 40) = 15.10, *p* = .001). Pair-wise comparisons with the Bonferroni correction revealed a significant difference between /v/ (*Md* = -.05, *SD* = .11) and /ɛ/ (*Md* = .26, *SD* = .21), *Z* = .54, *p* = .007. Segment /v/ was also significantly different from /oʊ/ (*Md* = .29, *SD* = 16), *Z* = .55, *p* = .005). No other pair-wise comparisons were significant.

The results suggested that both groups showed significant improvement from pre- to post-test in all target segments except /v/. D-group produced more intelligible segments on the pre-test than C-group, except for /oʊ/, in which case C-group had more intelligible productions. Within each group, there was considerable variation in the magnitude of change from pre- to post-test. Although no significant differences were found in the degree of intelligibility improvement between the two groups, the overall magnitude of change was higher in C-group. For /p/, the two groups showed a significant improvement from pre- to post-test, and the magnitude of change on average was comparable. For /v/, the two groups had highly intelligible productions on the pre-test, and while D-group's performance was quite higher, C-group had a greater magnitude of change.

In contrast, the two groups produced less intelligible target vowels on the pre-test compared to the target consonants. Although statistical analysis showed no significant differences in the intelligibility improvement of /ɛ/ between groups, C-group had a higher degree of improvement. For /oʊ/, the two groups showed a comparable magnitude of change, with C-group exhibiting only a slight change. It could be concluded that shadowing contributed to intelligibility improvement, regardless of group.

Most trainees in both groups, on average, revealed a considerable change on the post-test, which could not be interpreted by looking only at the degree of perceived convergence as judged by native English speakers. Trainees were mainly assigned to the groups based on their scores in the perceptual similarity tasks. Since perceived phonetic convergence did not appear to explain the improvement in intelligibility, a subsequent analysis assessed the relationship between intelligibility improvement and acoustic convergence.

4.5.5 Relationship between Phonetic Convergence and Magnitude of Intelligibility Improvement

The previous analysis failed to show significant differences between groups in terms of intelligibility improvement. Therefore, it was necessary to examine whether the degree of phonetic convergence in acoustic attributes and preexisting phonetic distance could explain variability in the magnitude of improvement regardless of group. To accomplish this, GEE models with an identity link function were used. This type of analysis is appropriate for repeated observations when data for each participant are correlated. The changes from pre- to post-test obtained from both groups were combined across the 12 target words, yielding 12 data points for each trainee. The dependent variable was the magnitude of change (i.e., the difference between the pre- and post-tests. The predictors were the four acoustic DID measures (VOT, vowel duration, F0, and F1*F2) and the baseline distance (or preexisting distance) of these measures. The proportions of perceived phonetic convergence were added to the model as another predictor to see whether it could explain intelligibility improvement regardless of group. To conduct this analysis, the baseline distances and DID measures were first converted to *z*-scores as these measurements were obtained from different scales. Note that the baseline distance and DID values in this analysis represented the values measured between the trainees and their trainers.

The GEE model results revealed a significant relationship between vowel duration DID values and magnitude of change in trainee intelligibility (B = .137, *p* = .026). F1*F2 DID values were also a significant predictor of the magnitude of change (B = .063, *p* = .016). These results suggested a direct relationship between trainees' convergence in vowel duration and F1*F2 towards their trainers and the degree of intelligibility improvement they exhibited after training.

The GEE models showed a significant relationship between the vowel duration baseline distance and degree of intelligibility improvement (B = -.101, *p* = .028). The F1*F2 baseline distance was also a significant predictor (B = -.135, *p* < .001). The negative beta indicated that the relationship was inverse. That is, the larger the vowel duration and F1*F2 baseline distances between trainer and trainee were, the smaller the effect of training was found to be. No other DID values or baseline distances were significant predictors of performance.

As mentioned earlier, dividing trainees into two groups based on their perceived phonetic convergence did not appear to affect segmental intelligibility. This observation was confirmed by the GEE results that perceived phonetic convergence was not a significant predictor of improvement. The results are summarized in Table 35.

Table 35. *Results from GEE Models with Independence Correlation Structure*

Parameter	B	Std. Error	95% Wald Confidence Interval		Wald Chi-Square	Sig.
			Lower	Upper		
(Intercept)	.227	.025	.18	.28	80.922	.000
*z*VOT DID	.011	.031	-.05	.07	.136	.712
*z*V-duration DID	.137	.062	.02	.26	4.959	.026*
*z*F0 DID	-.003	.037	-.08	.07	.005	.941
*z*F1*F2 DID	.063	.026	.01	.11	5.796	.016*
*z*VOT Baseline	.036	.035	-.03	.10	1.043	.307
*z*V-duration baseline	-.101	.046	-.19	-.01	4.832	.028*
*z*F0 baseline	.047	.037	-.03	.12	1.637	.201
*z*F1*F2 baseline	-.135	.035	-.20	-.07	14.717	<.001*
PC	-.062	.045	-.15	.03	1.957	.162

Note. PC stands for perceived convergence as assessed by native English speakers. All acoustic measures were converted to *z*-scores. Significant results are marked by an asterisk.

How native English listeners perceived trainees as more similar to or different from their trainers could not explain the magnitude of change exhibited after training. However, overall gains made by trainees after training in terms of segmental intelligibility, regardless of group, were predicted by acoustic convergence and the preexisting distance of some phonetic attributes. The results suggested that the more trainees converged their vowel duration and formants to their trainers, the better their performance was, and the further trainees were from their trainers in terms of vowel duration and F1*F2 before training, the worse their improvement was found to be. Thus, trainees whose vowel duration and formant baseline distances were closer to those of their trainers improved their segmental intelligibility more.

CHAPTER 5: DISCUSSION AND CONCLUSION

5.1 Discussion and Conclusion

This chapter provides a summary of the main findings and discusses them in light of the three research questions and previous studies. It also presents potential implications, limitations, and directions for future research.

5.2 Overview of the Study

The present work investigated phonetic convergence in L2 settings and its relevance to the acquisition of L2 segments. Specifically, it examined whether the degree of phonetic convergence that Arabic speakers showed towards native English speakers had any role in the improvement of segmental intelligibility after being trained by these speakers. The study also asked whether the preexisting phonetic distance between these learners and native speakers would determine the degree of phonetic convergence and magnitude of change they showed after training. Another aim was to examine whether shadowing was an effective way to improve segmental intelligibility in general or was constrained by learners' ability to phonetically converge to their trainers.

To accomplish this goal, the study went through several experimental phases. First, phonetic convergence was explored in learners' productions after they were exposed to five native English model talkers in non-interactive settings with numerous measurements. Following previous studies (e.g., Babel & Bulatov, 2012; Goldinger, 1998; Kim, 2013; Kim et al., 2011; Namy et al., 2002; Pardo, 2006; Pardo et al., 2017; Shockley et al., 2004), this study included XAB perceptual similarity judgments by native English listeners and acoustic phonetic measurements. Based mainly on the perceptual measures of phonetic convergence, Arabic-speaking L2 English learners were assigned to two groups based on the degree of convergence they exhibited to the model talkers. One group (C-group) received phonetic production training from the model talkers to whom they showed the highest degree of phonetic convergence, whereas the other (D-group) received training from the native model talkers they showed divergence from or the least convergence to. The criteria to match trainees with their trainers were largely centered on the perceptual measures of phonetic convergence. Nonetheless, acoustic measures were considered when a trainer received identical or nearly identical perceptual judgments with multiple model talkers. In this case, comparing convergence in acoustic attributes helped determine which model talker a trainee converged more to or diverged more from.

This study is one of the first to explore the role phonetic convergence towards native speakers plays in improving L2 pronunciation, particularly segmental intelligibility. The study used degree of convergence to train L2 learners to improve their pronunciation of some English sounds known to be difficult to Arabic speakers. This innovative methodology could pave the way for more studies on how L2 learners' ability to phonetically converge in some acoustic measures to native speakers in an L2 context might facilitate or accelerate their acquisition of L2 phonetics and phonology.

Another novel contribution is how thoroughly this study examined the phenomenon among Arabic speakers. No other research, to my knowledge, has examined phonetic convergence in Arabic speakers' L2 English speech. The study explored the degree of phonetic convergence these learners showed towards native English speakers using perceptual judgments and acoustic measures. In contrast, previous studies have investigated phonetic convergence in the speech of L2 learners who spoke languages other than Arabic, such as Mandarin (e.g., Olmstead et al., 2021), Korean (e.g., Hosseini-Kivanani et al., 2019; Kim, 2009; Kim et al., 2011; Kwon, 2021; Tobin, 2013), Polish (Rojczyk, 2012; Zając & Rojczyk, 2014), Spanish (Tobin, 2013, Ulbrich, 2021), Chinese (Ghanem, 2017), and French (Gessinger et al., 2020). The current study also asked whether the degree of acoustic convergence L2 learners showed to model talkers could be explained by preexisting phonetic distance (i.e., the baseline distance of measured acoustic features). The relationship between convergence in acoustic measures and perceived convergence as judged by native English speakers was examined as well. I used GEE modeling, controlling for model talker, to capture the ways in which these variables affected acoustic convergence. Therefore, this study delved deeply into the patterns of phonetic convergence exhibited by Arabic-speaking L2 learners of English.

As mentioned above, the assignment of trainees to C-group and D-group was based mostly on perceived convergence. The two groups received three consecutive days of production training on difficult segments in English nonsense words. They were trained using the shadowing technique under a low-variability paradigm in which each trainee received training from one model talker only. Pre- and post-productions of the same nonsense words were obtained to assess the effectiveness of the training. Pronunciation improvement was assessed by native English listeners judging segmental intelligibility.

Several parametric and non-parametric statistical tests were carried out to compare the segmental intelligibility gains made by the two groups. Additional analysis using GEE modeling was applied to capture the ways in which acoustic convergence and preexisting phonetic distance influenced the magnitude of their improvement in segmental intelligibility. The contribution of perceived convergence, as assessed by native English listeners, to segmental intelligibility improvement was evaluated as well. The following sections discuss the main findings in relation to the three research questions followed by a discussion of other notable findings.

5.3 Phonetic Convergence and Improved Segmental Intelligibility

The first research question asked, “Do Arabic learners of English improve their segmental productions more when they are trained by a native model talker to whom they phonetically converge?” To answer this question, experimental studies were devised to determine the impact of phonetic convergence on the pronunciation of four English segments (i.e., /p, v, ɛ, oʊ/). L2 learners were divided into two training groups based mainly on the perceived convergence assessed by native English speakers. The four target segments were presented to trainees in nonsense words, and their intelligibility improvement with these segments was likewise judged by native English speakers. The averaged intelligibility assessments indicated that both groups showed significant improvement from pre-test to post-test. No significant differences, however, were found between groups in terms of overall magnitude of change. A similar pattern was observed with the separate analysis of intelligibility improvement of the target segments. Both groups exhibited greater improvement from pre- to post-test in three target segments, /p, ɛ, oʊ/, while no improvement was found for /v/ due to the high performance of both groups before training leaving little room for improvement.

The majority of trainees regardless of group displayed substantial change on the post-test but with considerable variation. The expectation was that C-group would outperform D-group in segmental intelligibility, but group type (i.e., C-group and D-group) did not result in any significant difference in performance. It appears that assigning trainees to convergence or divergence groups based on how they were perceived as more similar to or different from their trainers did not explain the training results or the variation in their performance. This does not mean, however, that phonetic convergence plays no role in the acquisition of L2 pronunciation. It might be misleading to rely only on perceived convergence in capturing the ways phonetic convergence affects the acquisition of L2 segments. Therefore, it was important to scrutinize whether there was a relationship between acoustic convergence and the magnitude of segmental intelligibility improvement. If the same results were found, it would then be safer to conclude that phonetic convergence had no influence.

GEE modeling revealed that some patterns of acoustic convergence towards trainers, regardless of group, predicted trainees' overall segmental intelligibility gains. The findings suggested that the more trainees converged their vowel duration and formants to their trainers, the better their performance was. However, the degree of convergence in VOT and F0 was not a significant predictor of the overall change exhibited by trainees.

Another finding was that learners' overall performance differed according to the preexisting phonetic distance of some acoustic features between them and their trainers. At a featural level, the magnitude of the baseline distance (trainee's baseline production minus trainer's production) impacted performance. Learners showed greater improvement when they received training from model talkers whose vowel duration and formant frequencies were more similar to their own. A substantial inverse relationship was found between magnitude of change and preexisting phonetic distance only in vowel duration and vowel formants. However, no significant relationship was found between preexisting phonetic distance of VOT and F0 and the magnitude of change. Accordingly, the farther away trainees were from their trainer in terms of vowel duration and formant frequencies, the less improvement they achieved.

Some findings regarding preexisting distance partially align with those reported in Probst at el. (2002), though their study was limited to how the auxiliary verb was stressed in sentences and no acoustic measurements were used. In their study, participants who were matched with native speakers with the most similar speech rate showed more improvement. In a similar way, this study found that the closer trainees were in terms of vowel duration to the native speakers they shadowed, the greater their segmental intelligibility improvement was.

The question here is why phonetic convergence in vowel duration and vowel spectra were significantly related to overall improvement rather than convergence in VOT and F0. A likely explanation is that the GEE model included the overall intelligibility assessment of all target segments. Each target segment was presented in three nonsense words. Thus, there were 12 data points for each trainee, which included the examined consonants and vowels, and 50% of segments evaluated by listeners were vowels. Vowel duration and formant frequencies were evidently more related to vowel properties than to consonants. It is likely that convergence in vowel duration and vowel spectra was a significant predictor of trainees' overall performance, presumably due to the larger number of assessed segments that were vowels.

Another issue is that VOT patterns shown by learners in their convergence or preexisting phonetic distance did not explain overall improvement. VOT is one of the main acoustic properties used to categorize fortis and lenis stops in English but was relevant to only 25% of target segments (i.e., /p/). Thus, it was not surprising that degree of VOT convergence and preexisting distance were not associated with any segmental intelligibility improvement. This study could not make generalizations about the acquisition of English stop VOT due to the small number of words targeting stops, and the study did not have the chance to do acoustic measurements to assess the efficacy of training. However, different findings might come from examining the actual measurements of VOT and comparing them to the VOT of English stops produced by native speakers.

The pronunciation gains achieved by learners were not explained by F0 convergence or preexisting distance. F0 has a strong correlation with gender, with female speakers having a higher F0 than male speakers. This study investigated phonetic convergence patterns only in the speech of female Arabic speakers shadowing female English speakers to minimize cross-gender differences. Therefore, it was not surprising to find less convergence (or even maintenance) in F0 values exhibited by those learners to their model talkers. This might be attributed to the preexisting F0 distance between learners and the model talkers they shadowed. If trainees were matched with trainers of the opposite gender, greater convergence might have been seen in the F0. However, it is hard to predict a specific pattern since previous studies have shown inconsistent results regarding F0 convergence across gender. For example, Babel and Bulatov (2012) found that male participants converged their F0 more to a male model talker than female participants, where the original recordings were used without manipulation. Their findings suggested that F0 convergence might be inhibited in cross-gender conditions. However, Hosseini-Kivanani et al. (2019) showed that male shadowers converged their F0 to a female model talker more than female shadowers, which suggested that having different genders resulted in more F0 convergence.

In this study, degree of F0 convergence towards native English speakers appeared to have no association with the magnitude of segmental intelligibility improvement displayed by trainees. A possible reason is that F0 is not a primary cue to determine vowel quality in English, as reported by previous studies (e.g., Delattre et al., 1952; Irino & Patterson, 2002; Raphael, 2021; Smith & Patterson, 2005). That could explain this study finding no association between degree of convergence or preexisting distance in F0 and the magnitude of segmental intelligibility improvement.

In contrast, convergence in F0 is likely to be more relevant to pronunciation improvement with suprasegmental structures at word-level or sentence-level prosody. For example, F0 is one of the main correlates of stressed syllables in languages such as English; therefore, L2 learners of English might benefit from the ability to converge their F0 to learn the patterns of stress assignment. More research is thus needed to evaluate the influence of F0 convergence on learning L2 suprasegmental structures.

Another possibility is that convergence in F0 might be determined by language-specific phonetics. Native speakers of tonal languages (e.g., Mandarin) and pitch-accent languages (e.g., Japanese) might exhibit different patterns of F0 convergence from those by Arabic-speaking L2 learners in this study. For example, Mandarin uses suprasegmental cues such as F0 that serve crucial lexical functions (Lin et al., 2014; Qin et al., 2016). Another example is Seoul Korean speakers who use post-stop F0 as an important acoustic cue to distinguish aspirated and lenis stops (Kang & Guion, 2006; Kim et al., 2002; Kim & Tremblay 2021; Kong et al., 2011; Silva, 2006). Thus, more convergence in F0 might be displayed by native speakers of these languages, which could be correlated with L2 segmental and suprasegmental learning. Alternatively, L2 learners of these languages who show more convergence in F0 towards their trainers might display more pronunciation improvement.

The GEE models on the relationship between acoustic measures and perceived phonetic convergence revealed a substantial relationship between acoustic convergence in VOT, F0, and F1*F2 and the degree of perceived convergence. The more learners converged these acoustic features to model talkers, the more they were judged as similar to the model talkers. However, convergence in vowel duration did not contribute to how native English listeners perceived the similarity between learners and model talkers, though vowel duration was the only acoustic feature that had significant average convergence. These findings are inconsistent with studies that explored the relationship between acoustic measures and perceived convergence in L1 settings. Pardo et al. (2013), for example, reported that variation in the degree of convergence in vowel duration was the strongest predictor of listeners' assessment of pronunciation similarity. This inconsistency might stem from listeners' judgments being based on L2 productions rather than native productions, as judgments about pronunciation similarity could be affected by whether the shadowers are native or non-native speakers. It might be that L1 judges did not fully register how L2 speakers converged their vowel duration to native speakers, with other factors affecting their perceptions, such as degree of foreign accent and comprehensibility. The researcher found no studies on the relationship between acoustic measures and perceived convergence in L2 speech. Therefore, it would be interesting to examine this relationship with L2 learners from different linguistic backgrounds to see whether these findings would be supported.

The degree of convergence in vowel duration and formants found to be a significant predictor of segmental intelligibility improvement. However, vowel duration did not contribute to the pronunciation similarity judgements that were used to match trainees with their trainers. This might explain the failure to find any significant differences between the achieved improvement of the two groups. The assignment of trainees to two groups was determined mainly by their degree of perceived convergence, which was not influenced by vowel duration convergence.

To answer the first research question, L2 learners trained by English speakers to whom they exhibited more phonetic convergence appeared to improve their segmental intelligibility even after a short period of production training. However, this depends on what convergence is being looked at. Convergence in L2 learners' speech as judged by native English speakers did not explain the magnitude of improvement. Rather, acoustic convergence in acoustic measures, particularly in vowel duration and vowel spectra, resulted in more segmental intelligibility improvement.

5.4 Relationship Between Preexisting Acoustic Distance and Degree of Convergence

The second research question asked, "Does the preexisting phonetic distance between native model talkers and Arabic learners of English determine the degree of convergence?" As expected, GEE analyses controlling for model talker suggested the magnitude of preexisting phonetic distance explained the degree of convergence in the four measured acoustic features (i.e., VOT, vowel duration, F0, and F1*F2). A direct relationship was found between how far L2 learners were from model talkers and degree of acoustic convergence. In other words, the larger the preexisting distance in a given phonetic dimension was, the greater the degree of convergence. It could be inferred that phonetic convergence increased with a phonetic feature when the preexisting phonetic distance was larger as L2 learners had more room to converge. Kim (2012) had similar findings in an L1 setting. Her study was conducted on the speech of native English speakers shadowing model talkers with three different linguistic distances: a native English model talker with the same L1 dialect, a native English model talker with the same L1 but a different dialect, and a high-proficiency non-native model talker. The degree of phonetic convergence was influenced by the preexisting phonetic distance, regardless of the language distance between interlocutors. Although Kim used a formula to measure phonetic convergence and a statistical analysis different from the one in this study, she obtained basically the same findings regarding preexisting phonetic distance.

This finding might be due to the methodology. DID has been the most popular way to evaluate phonetic convergence, but recent papers (e.g., Cohen Priva & Sanker, 2019; MacLeod, 2021) have identified problems with these measurements. For instance, MacLeod (2021) argued that many studies have reported a direct relationship between DID and baseline distance, that is, the larger the baseline, the larger the DID. Thus, the farther participants are from the model talker, the more convergence they will show. But this does not reflect actual convergence. DID can capture broad convergence when the baseline is really different from the model talker, but problems appear when the baseline is small or close to the model talker. First, the relationship between DID and baseline is always biased due to the way they are calculated (DID = absolute baseline distance – shadowing distance). This means researchers will always find a direct relationship between the baseline distance and DID. Another issue with DID, as stated by MacLeod (2021), is that when the baseline distance was very small, divergence was found, even if participants converged and the model actually had an effect on their speech if it was measured by linear combination method. DID also does not distinguish between no change and overconvergence. No change means participant's absolute baseline distance is similar to shadowed distance (i.e., baseline distance = 1 and shadowed distance =1, then DID = 0). Overconvergence means a participant converge to a model talker to the extent that surpasses model talker's value. For example, if the baseline value = 3 and model value = 4, the absolute baseline distance = 1; and the shadowed value = 5, then the absolute shadowed value = 1. The DID = 0, which the same value when no change happens.

MacLeod (2021) employed two different methods to measure convergence in the same data: DID and linear combination. In DID, the independent variable was absolute baseline distance, and the dependent variable was DID. Baseline distance was a significant predictor of the degree of convergence (the larger the baseline, the larger the DID). However, when linear combination was used, this direct relationship was not found. In linear combination, the dependent variable was the actual shadowed value (not the absolute distance), and the independent variables were actual baseline value, model talker's value, and the interaction between model talker's value and absolute baseline distance. This method compared the baseline, model, and shadowed values of the same word rather than using the averages, as many studies have done. Convergence was detected when the model value affected the shadowed value. Although that study used the same data, the analysis of linear combination did not show any significant effect from baseline distance on convergence. This was consistent with Cohen Priva and Sanker (2019), who showed that measuring DID to find convergence carried some bias, but their study was conducted on speech from real conversations, and they took measurements at different levels of the conversation. They suggested that linear combination appeared to be a better alternative to measure phonetic convergence.

However, if linear combination is used to assess phonetic convergence, convergence cannot be compared to other variables. This is because there is no value for convergence that can be a predictor for other variables like perceived convergence (AXB) or the magnitude of change after training as examined in the current study. If linear combination had been used in this study, it would be impossible to compare the degree of acoustic convergence to perceived convergence or to the improvement of segmental intelligibility. It would, however, be interesting to analyze the data in this study with linear combination to see whether baseline distance would still be a significant predictor of phonetic convergence.

5.5 The Effect of Shadowing on Segmental Intelligibility Improvement

The last research question asked, “Is the shadowing technique generally effective in improving pronunciation, or is it constrained by L2 learners’ degree of phonetic convergence?” The production training revealed that L2 learners in both groups improved their segmental intelligibility substantially between the pre-test and post-test. The results suggested that shadowing could substantially promote the segmental intelligibility of Arabic-speaking L2 learners of English. This finding was not surprising, as previous studies on speaking and other L2 skills (e.g., Bovee & Stewart, 2009; Foote & McDonough, 2017; Hamada, 2016; Horiyama, 2012; Hsieh et al., 2013; Kadota, 2019; Wang, 2017) have reported empirical support for the efficacy of shadowing in L2 learning. While the findings of the current study suggested that shadowing was an effective tool in pronunciation improvement, its effectiveness might be increased by the degree of convergence in vowel duration and vowel spectra that trainees showed towards their trainers. Another factor that likely enhanced the effectiveness of shadowing was the small preexisting distance in vowel duration and vowel spectra. That is, shadowing native speakers whose voices were similar to those of L2 learners in terms of vowel duration and vowel spectra resulted in more improvement.

These two findings might seem contradictory. Based on the DID metric, the current study found baseline distance was a significant predictor of convergence in acoustic attributes: that is, the larger the baseline distance, the larger the convergence. However, this might not reflect the real patterns of convergence, as discussed in the answer to the second research question. A small distance might not entail that participants could not converge to the model. Statistical analysis showed that convergence (DID) in vowel duration and vowel spectra were significant predictors of trainees' performance; at the same time, there was an inverse relationship between the baseline distance of these features and trainee performance: that is, the larger the baseline, the less improvement trainees showed. A plausible explanation is that the key to benefit more from trainers is being close to them, either in baseline distance or from converging to them.

5.6 Other Notable Findings

Another interesting finding from the perceptual similarity judgments was that Model Talker 3 evoked more perceived convergence than the other model talkers, although she was only significantly higher from Model Talkers 2 and 5. Trainees' shadowed speech was perceived as 70% similar to that of Model Talker 3. The higher degree of perceived convergence in shadowing Model Talker 3 made her the trainer for six trainees in C-group. Therefore, it would be interesting to examine what characteristics she had that could have contributed to this result. Examining convergence in the acoustic features demonstrated some striking patterns. Since vowel duration was not a significant predictor of perceived convergence, it was excluded from the analysis. The average VOT, F0, and F1*F2 DID values between each model talker and all trainees are presented in Table 36.

Table 36. *Average DID Values across the Five Model Talkers*

Model Talker	VOT	F0	F1*F2
1	-.27	3.82	19.70
2	-2.45	5.31	23.27
3	-.26	2.54	34.04
4	-1.13	3.61	10.81
5	-1.34	-1.25	-12.11

The average VOT and F0 DID values showed no large differences across model talkers, although Model Talker 3 evoked the highest convergence with F1*F2 DID. Accordingly, it would be worth looking at the patterns of convergence in the individual shadowed vowels. Chapter 3 provided a brief description of the general tendencies of how trainees converged individual vowels towards the model talkers. However, it was beyond the scope of this study to examine the convergence patterns exhibited by individual vowels.

The overall observation was that /æ/ was imitated by trainees more than the other vowels with Model Talker 3. Compared to the other model talkers, Model Talker 3 produced /æ/ with the highest F2 and lowest F1. Although the five model talkers reported speaking a Midwest dialect and were all born and raised in Wisconsin, /æ/ raising was most evident in the words spoken by Model Talker 3, particularly the words *tagger* and *dagger*. What might lead to more vowel raising in these words was the environment in which /æ/ appeared. According to some studies on North American vowel shifts (e.g., Bauer & Parker, 2008; Benson et al., 2011; Purnell, 2009; Zeller, 1997), speakers of the affected dialects raise /æ/ more when it occurs before velar consonants, referred to as "prevelar raising." In the baseline productions, all L2 learners produced *tagger* and *dagger* without any raising, but all of them except one pronounced them with an obvious raising in the shadowing tasks. It appears the learners were not aware of the /æ/ raising. Only one reported the accent of that model talker as sounding different and produced those words without raising. Thus, the high degree of perceived convergence towards Model Talker 3 appeared to originate from the imitation of /æ/.

Another interesting finding was that L2 learners produced a more intelligible /v/ even before training. Given that the Arabic phonemic inventory does not have /p/ and /v/, these segments were expected to be problematic for Arabic L2 learners of English. This study did not use acoustic measures to evaluate segmental improvement in trainees' productions; therefore, it is possible they produced /p/ without initial aspiration on the pre-test, which could have influenced how listeners perceived the segment.

In a study on the acquisition of the English /p/-/b/ contrast by native Arabic speakers, Eckman et al. (2015) found some learners produced significant VOT distinctions between the two sounds, but this contrast was not perceived by the English transcribers. Furthermore, in cross-language perception studies (e.g., Lotz et al., 1960), initial aspiration was found to be a more prominent cue in detecting fortis stops among English listeners than absence of voicing. Therefore, /p/ lacking aspiration in the pre-test was likely perceived as /b/. After training, trainees apparently produced an aspirated /p/ with a more native-like VOT, which resulted in more correct identifications of /p/. This could be demonstrated empirically in future research by looking at the acoustic analysis of VOT patterns exhibited by trainees.

In addition, trainees in both groups produced more intelligible consonants than vowels before training. This result is consistent with Alshangiti (2015), who found that English vowels were more problematic than consonants for Saudi L2 English learners. She attributed this difficulty to the relative sizes of the sound inventories of Arabic and English. Arabic has 28 consonants, and English has a similar amount (24), making it easier to map English consonants to Arabic ones. However, Arabic only has six vowels compared to the 17 in British English, making these sounds more challenging.

The two groups showed a substantial difference from pre- to post-test in their intelligibility improvement of /ɛ/ and /oʊ/. Although C-group showed a slightly higher magnitude of change, statistical analysis failed to find any significant differences between groups in terms of magnitude of change. A potential cause is sample size, as 10 trainees in each group might not have been large enough to reveal statistical significance.

5.7 Limitations, Implications, and Directions for Future Studies

Similar to many studies on phonetic convergence (e.g., Babel, 2009, 2010; Gessinger et al., 2021; Gnevsheva et al., 2021; Goldinger, 1998; Goldinger & Azuma, 2004; Hosseini-Kivanani et al., 2019; Kim, 2012; Kwon, 2019, 2021; Namy et al., 2002; Nielsen, 2011; Tobin, 2013), the current study measured patterns of convergence in non-interactive settings, with participants passively exposed to the model talkers. However, findings based on laboratory settings can minimize the difficulties encountered in actual L2 communication. It should also be noted that phonetic convergence was explored in English words lacking segments that would pose difficulties for Arabic speakers learning English. Therefore, the degree and patterns of convergence learners showed to native English speakers might be altered in real interactive settings or when phonetic convergence is assessed in segmental structures that do not exist in their native language. Nevertheless, it would be interesting for future investigations to elicit phonetic convergence exhibited by Arabic speakers in a more natural interactive setting in terms of various segmental and suprasegmental structures. This would provide a broader understanding of this phenomenon and the role of L1 phonetics and phonology in convergence.

Similar to numerous studies (e.g., Babel & Bulatov, 2012; Pardo et al., 2013; Pardo et al., 2017), learners in the current study showed convergence with some features and divergence with others. For instance, some showed convergence in vowel spectra and divergence in VOT and F0, while others showed convergence in VOT and vowel duration but no change in other features. Furthermore, variation in convergence patterns was found within the same participant, who converged one acoustic feature with one model talker but diverged the same feature with another model talker. These findings are inconsistent with the Interactive Automatic Alignment Account of dialogue (Pickering & Garrod, 2004), which proposes convergence to be an automatic process that occurs subconsciously. Nevertheless, the patterns of phonetic convergence in this study could not be attributed to an automatic process that occurs at all times.

This study adds to the field of speech perception by offering insights into the extent to which speech perception affects speech production in L2 settings. The findings suggested the L2 learners, though with considerable variation, could phonetically converge to native speakers. This in turn suggested learners could hear L2 phonetic details and consequently alter their speech after being exposed to native speakers. These findings support speech perception models positing a strong link between perception and production, such as Motor Theory by Liberman and Mattingly (1985) and exemplar-based models by Johnson (1997, 2006) and Pierrehumbert (2001, 2003).

This study extensively explored the relationship between phonetic convergence and L2 pronunciation learning. Arabic L2 learners of English showed phonetic convergence with some acoustic features, particularly vowel duration and vowel spectra, to improve their segmental intelligibility. The empirical evidence indicated learners used some aspects of phonetic convergence as a driving mechanism to acquire L2 segmental structures. This would appear to support Nguyen and Delvaux's (2015) proposal that, based on studies on phonetic drift, phonetic convergence might be used to develop the L2 phonetic and phonological system. The current study found a direct relationship between phonetic convergence in vowel duration and vowel spectra and the magnitude of segmental intelligibility improvement. However, it was unclear which type of convergence would result in more gains in L2 pronunciation: the ability to converge vowel duration or vowel spectra. These results thus lay the foundation for future studies to investigate whether trainees would benefit more from convergence in vowel duration or vowel spectra at the segmental or suprasegmental level. In addition, trainees whose vowel duration and vowel spectra were more similar to their trainers improved their segmental intelligibility more than those who were dissimilar to their trainers. Thus, it would be interesting in future studies to examine which is more effective in L2 pronunciation improvement, L2 learners' ability to converge vowel duration and vowel spectra to native speakers or having similar vowel duration and vowel spectra from the outset.

A major limitation of this study had to do with sampling. It would have been more ideal to recruit participants whose English proficiency, years of residency, age, and language experience were as comparable as possible, but it was challenging to find enough participants with similar levels during the pandemic. As a result, it was impossible to control for these factors in this study. Another limitation was that the small sample made it impossible to eliminate trainees who had high intelligibility scores on the pre-test. D-group was generally perceived as more intelligible before the training. Not only did D-group perform better overall than C-group on the pre-test, but their average age was lower than that of C-group. This observation was confirmed by an independent-sample *t*-test, which demonstrated that the age of C-group (M = 31.1, SD = 6.33) was significantly higher than D-group (M = 25.6, SD = 4.43), $t(18)$ =2.13, p < .05. These limitations should be addressed in future studies by controlling more for these factors. As noted earlier, all trainees and model talkers in this study were women. Many previous studies on how social factors affect phonetic convergence have reported that the model talker's gender affected the likelihood of convergence. However, these studies showed conflicting results. In some (e.g., Namy et al., 2002), women converged more than men, whereas others (e.g., Pardo,

2006) found male speakers showed higher convergence, and some failed to detect any differences in convergence patterns according to gender (e.g., Thomson et al., 2001). Since this study analyzed the convergence of female trainees shadowing female model talkers, the findings cannot be generalized to the convergence patterns displayed by male speakers in same-gender or mixed-gender shadowing tasks. How gender influences the likelihood of phonetic convergence shown by Arabic speakers should thus be examined in future studies.

There is ample evidence of the effectiveness of high-variability training in improving the speech perception of L2 learners, who not only perform better after training but can also make generalizations to untrained material and display long-term retention of what they have learned. Yet the focus of this study was on the relationship between phonetic convergence and the acquisition of L2 segments. Consequently, this study employed low-variability training, with each trainee having only one model talker. One motive for this was to lessen the impact of model talker variability as some studies have reported that shadowing several model talkers in the same block decreased the occurrence of phonetic convergence (Bable & McGuire, 2015). Hence, the high-variability paradigm that involves exposing L2 learners to multiple native speakers could hinder the ability of learners to converge to the model talkers. Accordingly, it might be impossible to draw reliable conclusions about the role of phonetic convergence in learning L2 pronunciation, but high-variability training has been used more extensively in L2 speech perception studies. Though some studies (e.g., Bradlow et al., 1997) have reported that learners showed improvement in their L2 productions after perceptual training, their main aim was to improve the perception of L2 categories.

This study found low-variability production training to be effective at improving L2 learners' segmental intelligibility. These findings are applicable only to short-term gains, however, and cannot be generalized to other aspects of learning since the study did not test learners' ability to generalize what was learned to untrained materials or their ability to fully retain categories in a follow-up test. Further investigation is therefore needed to determine the extent to which low-variability training affects generalization and long-term retention.

In the literature, it remains unclear how variability in L2 production training can lead to outcomes similar to that of high-variability speech perception training. In the current study, L2 learners were trained using a low-variability paradigm. Specifically, they were trained to produce the target segments in the same phonological environment in monosyllabic words. This method was effective at improving segmental intelligibility at least in the short term. However, it could be feasible to incorporate high variability in production training while keeping the trainer-trainee convergence effect. That is, a future study might expose L2 learners to one native speaker but present the target segment in several phonetic contexts in words with different syllable structures. That would exploit the advantages of high-variability training while sustaining the benefits of phonetic convergence.

5.8 Conclusions

This study investigated the role of phonetic convergence in the acquisition of L2 segments. The first aim was to determine how the phonetic convergence Arabic speakers showed towards native English speakers might affect their improvement of segmental intelligibility after being trained by those speakers. This aim was addressed by exploring phonetic convergence exhibited by Arabic L2 learners of English towards five native English speakers using acoustic measures and perceptual similarity judgments of native English listeners. Learners were then assigned to two training groups based largely on the perceptual similarity judgements under a low-variability paradigm. One group was trained by the model talkers to whom they exhibited the highest degree of convergence, while the other group received training from the model talkers they showed divergence from or the least convergence to. This criterion for assigning trainees did not explain the magnitude of segmental intelligibility improvement. Nonetheless, the degree of convergence in some acoustic measures (i.e., vowel duration and formants), regardless of group, explained the overall segmental intelligibility gains made by the trainees. Thus, the more trainees converged their vowel duration and vowel spectra to their trainers, the greater improvement they achieved. These results could be used to create a baseline for more research on the influence of phonetic convergence in L2 speech acquisition.

The second aim was to determine whether a relationship between the preexisting phonetic distance between L2 learners and native speakers would predict the degree of learners' phonetic convergence and improvement. The data showed the degree of phonetic convergence in acoustic features was correlated with the preexisting phonetic distance. Phonetic convergence increased in a single phonetic feature when the preexisting phonetic distance was larger. However, this finding might not reflect the real patterns of convergence exhibited by learners since the relationship between DID and baseline is always biased due to the way it is calculated.

The final aim of the study was to explore the effect of shadowing on segmental intelligibility in general and whether its effectiveness would be enhanced by learners' ability to phonetically converge to their trainers. The findings suggested that shadowing was an effective tool promoting L2 segmental intelligibility, which might be enhanced by the degree of convergence in vowel duration and vowel spectra that trainees showed towards their trainers. Moreover, similarity in vowel duration and spectra was found to increase the effectiveness of shadowing. With this in mind, shadowing appears to be an effective training tool in learning L2 segments, but its efficacy might be enhanced by having trainers whose vowel duration and spectra are similar to the learners or having trainers to whom the learners show greater convergence in this regard.

REFERENCES

Abrego-Collier, C., Grove, J., Sonderegger, M., & Alan, C. L. (2011). Effects of Speaker Evaluation on Phonetic Convergence. *In ICPhS* (pp. 192-195).

Acton, W. (1984). Changing fossilized pronunciation. *TESOL quarterly, 18(1), 71-85*.

Alfaifi, A. A. (2019). *Syllabification of coda consonant clusters in Najdi and Hijazi Arabic*. [Doctoral dissertation, George Mason University].

Al-Khairy, M. A. (2005). *Acoustic characteristics of Arabic fricatives*. [Doctoral dissertation, University of Florida].

Alshangiti, W. M. M. (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* (Doctoral dissertation, UCL (University College London)).

Althubyani, R. & Park, H. (2019, October 4-5). The Effect of Dialects on Phonetic Convergence in Non-native Settings [Poster presentation]. The meeting of the 24th Annual Mid-Continental Phonetics & Phonology Conference, University of Wisconsin-Milwaukee, Milwaukee, WI, United States.

Armson, J., & Kalinowski, J. (1994). Interpreting results of the fluent speech paradigm in stuttering research: Difficulties in separating cause from effect. *Journal of Speech and Hearing Research, 37*, 69–82.

Babel, M. E. (2009). Phonetic and social selectivity in speech accommodation. [Doctoral dissertation, University of California, Berkeley].

Babel, M. E. (2010). Dialect divergence and convergence in New Zealand English. *Language in Society, 39*(04), 437-456.

Babel, M. E. (2012). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics, 40*(1), 177-189.

Babel, M., & Bulatov, D. (2012). The role of fundamental frequency in phonetic accommodation. *Language and Speech, 55*(2), 231-248.

Babel, M., & McGuire, G. (2015). The effects of talker variability on phonetic accommodation. In *ICPhS*.

Babel, M., McGuire, G., Walters, S., & Nicholls, A. (2014). Novelty and social preference in phonetic accommodation. Journal of the Association for Laboratory Phonology*, 5*(1), 123-150.

Barriuso, T. A. (2018). *The L2 Acquisition of Phonemes and Allophones under Various Exposure Conditions* (Doctoral dissertation, The University of Utah).

Barriuso, T. A., & Hayes-Harb, R. (2018). High variability phonetic training as a bridge from research to practice. *CATESOL Journal, 30*(1), 177-194.

Bauer, M., & Parker, F. (2008). /æ/-raising in Wisconsin English. *American Speech, 83*(4), 403-431.

Bell, A. (1984). Language style as audience design. *Language in Society*, *13*(2), 145-204.

Benson, E. J., Fox, M. J., & Balkman, J. (2011). The bag that Scott bought: The low vowels in northwest Wisconsin. *American Speech, 86*(3), 271-311.

Boersma, Paul & Weenink, David (2021). Praat: doing phonetics by computer [Computer program]. Version 6.2.02, retrieved 2 December 2021 from http://www.praat.org/

Bovee, N., & Stewart, J. (2009). The utility of shadowing. In A. M. Stoke (Ed.), JALT 2008 conference proceedings (pp. 888-900). Tokyo: JALT.

Bradlow, A. R. (2008). Training non-native language sound patterns: Lessons from training

Japanese adults on the English. *Phonology of Second Language Acquisition, 36*, 287-308.

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *The Journal of the Acoustical Society of America, 101*(4), 2299–2310. https://doi.org/10.1121/1.418276

Brysbaert, M. & New, B. (2009) Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41* (4), 977-990.

Bürki, A., Spinelli, E., & Gaskell, M. G. (2012). A written word is worth a thousand spoken words: The influence of spelling on spoken-word production. *Journal of Memory and Language*, *67*(4), 449-467.

Celce-Murcia, M., Brinton, D. M., & Goodwin, J. M. (1996). *Teaching pronunciation: A reference for teachers of English to speakers of other languages*. New York, NY: Cambridge University Press.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America*, *25*(5), 975-979.

Clark, H. H., & Murphy, G. L. (1982). Audience design in meaning and reference. *Advances in psychology, 9*, 287-299.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46.

Costa, A., Pickering, M. J., & Sorace, A. (2008). Alignment in second language dialogue. *Language and Cognitive Processes*, *23*(4), 528-556.

Crystal, T. H., & House, A. S. (1988a). A note on the durations of fricatives in American

English. *The Journal of the Acoustical Society of America*, *84*(5), 1932-1935.

Crystal, T. H., & House, A. S. (1988b). Segmental durations in connected-speech signals: Current results. *The Journal of the Acoustical Society of America*, *83*(4), 1553-1573.

Danchenko, N. (2011). Pronunciation and independent work: Embedding pronunciation into academic English skill classes. *Advances in Language and Literary Studies*, *2*(2), 171-184.

Davies, Mark. (2008-) *The Corpus of Contemporary American English (COCA): 560 million words, 1990-present*. Available online at https://corpus.byu.edu/coca/.

Delattre, P., Liberman, A. M., Cooper, F. S., & Gerstman, L. J. (1952). An experimental study of the acoustic determinants of vowel color; observations on one-and two-formant vowels synthesized from spectrographic patterns. *Word, 8*(3), 195-210.

Ding, S., Liberatore, C., Sonsaat, S., Lučić, I., Silpachai, A., Zhao, G., ... & Gutierrez-Osuna, R. (2019). Golden speaker builder–An interactive tool for pronunciation training. *Speech Communication*, *115*, 51-66.

Eckman, F., Iverson, G., & Song, J. (2015). Overt and covert contrast in L2 phonology. *Journal of Second Language Pronunciation, 1*(2), 254-278.

Edwards, J. G. H., & Zampini, M. L. (Eds.). (2008). *Phonology and second language acquisition* (Vol. 36). Philadelphia, PA: John Benjamins Publishing.

Evans, B. G., & Alshangiti, W. (2018). The perception and production of British English vowels and consonants by Arabic learners of English. *Journal of Phonetics*, *68*, 15-31.

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

Foote, J. A. (2015). *Pronunciation pedagogy and speech perception: Three studies*. [Doctoral dissertation, Concordia University].

Foote, J. A., & McDonough, K. (2017). Using shadowing with mobile technology to improve L2 pronunciation. *Journal of Second Language Pronunciation, 3*(1), 34-56.

Foster, & Cole, J. (2020). Effects of word-frequency and social evaluation on phonetic accommodation. *The Journal of the Acoustical Society of America, 148*(4), 2761–2761. https://doi.org/10.1121/1.5147678 <https://doi.org/10.1121/1.514767>

Gessinger, I., Möbius, B., Andreeva, B., Raveh, E., & Steiner, I. (2020). Phonetic Accommodation of L2 German Speakers to the Virtual Language Learning Tutor Mirabella. *INTERSPEECH*, 4118-4122.

Gessinger, I., Raveh, E., Steiner, I., & Möbius, B. (2021). Phonetic accommodation to natural and synthetic voices: Behavior of groups and individuals in speech shadowing. *Speech Communication, 127*, 43-63.

Ghanem, R. (2017). *Nonnative speakers' alignment of linguistic features with different interlocutors*. [Doctoral dissertation, Northern Arizona University].

Giles, H., Coupland, N., & Coupland, I. U. S. T. I. N. E. (1991). 1. Accommodation theory: Communication, context, and. *Contexts of accommodation: Developments in applied sociolinguistics*, *1*.

Gnevsheva, K., Szakay, A., & Jansen, S. (2021). Phonetic convergence across dialect boundaries in first and second language speakers. *Journal of Phonetics, 89*, Article 101110.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological review*, *105*(2), 251-279.

Goldinger, S. D., & Azuma, T. (2004). Episodic memory reflected in printed word naming. *Psychonomic bulletin & review*, *11*(4), 716-722.

Hamada, Y. (2016). Shadowing: Who benefits and how? Uncovering a booming EFL teaching technique for listening comprehension. *Language Teaching Research*, *20*(1), 35-52.

Hamada, Y. (2017). *Teaching EFL Learners Shadowing for Listening: Developing learners' bottom-up skills*. New, NY: Routledge.

Horiyama, A. (2012). The development of English language skills through shadowing exercises. *Journal of Bunkyo Gakuin University of Foreign Studies, Bunkyo Gakuin Junior College*, (12), 113-123.

Horton, W. S., & Gerrig, R. J. (2005). The impact of memory demands on audience design during language production. *Cognition, 96*(2), 127-142.

Hosseini-Kivanani, N., Tobin, S. J., & Gafos, A. I. (2019). Phonetic accommodation in the fundamental frequency of Korean-English bilinguals and English monolinguals.

Hsieh, K. T., Dong, D. H., & Wang, L. Y. (2013). A preliminary study of applying shadowing technique to English intonation instruction. *Taiwan Journal of Linguistics*, *11*(2), 43-65

Hwang, J., Brennan, S. E., & Huffman, M. K. (2015). Phonetic adaptation in non-native spoken dialogue: Effects of priming and audience design. *Journal of Memory and Language, 81*, 72–90. https://doi.org/10.1016/j.jml.2015.01.001

Irino, T., & Patterson, R. D. (2002). Segregating information about the size and shape of the vocal tract using a time-domain auditory model: The stabilised wavelet-Mellin transform. *Speech Communication, 36*(3-4), 181-203.

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.

Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. *The Journal of the Acoustical Society of America*, *108*(3), 1252-1263

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In Johnson & Mullennix (eds) *Talker Variability in Speech Processing.* San Diego: Academic Press. pp. 145-165.

Johnson, K. (2006). Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics 34*, 485-499.

Kadota, S. (2019). *Shadowing as a practice in second language acquisition: Connecting inputs and outputs*. New York: Routledge.

Kang, K. H., & Guion, S. G. (2006). Phonological systems in bilinguals: Age of learning effects on the stop consonant systems of Korean-English bilinguals. *The Journal of the Acoustical Society of America*, *119*(3), 1672-1683.

Kim, H., & Tremblay, A. (2021). Korean listeners' processing of suprasegmental lexical contrasts in Korean and English: A cue-based transfer approach. *Journal of Phonetics*, *87*, Article 101059.

Kim, M. (2012). *Phonetic accommodation after auditory exposure to native and nonnative speech.* ([Doctoral dissertation, Northwestern University].

Kim, M. (2009). Phonetic accommodation in conversations between native and non-native speakers. *The Journal of the Acoustical Society of America*, *125*(4), 2764-2764.

Kim, M., Horton, W. S., & Bradlow, A. R. (2011). Phonetic convergence in spontaneous conversations as a function of interlocutor language distance. Journal of the Association for Laboratory Phonology*, 2*(1), 125-156.

Kim, M. R., Beddor, P. S., & Horrocks, J. (2002). The contribution of consonantal and vocalic information to the perception of Korean initial stops. *Journal of Phonetics*, *30*(1), 77-100.

Kong, E. J., Beckman, M. E., & Edwards, J. (2011). Why are Korean tense stops acquired so early?: The role of acoustic properties. *Journal of phonetics*, *39*(2), 196-211.

Kwon, H. (2019). The role of native phonology in spontaneous imitation: Evidence from Seoul Korean. *Journal of the Association for Laboratory Phonology, 10*(1).

Kwon. (2021). A non-contrastive cue in spontaneous imitation: Comparing mono- and bilingual imitators. Journal of Phonetics, 88, 101083–. https://doi.org/10.1016/j.wocn.2021.101083

Ladefoged, P., & Johnson, K. (2014). *A course in phonetics*. Nelson Education.

Lambert, S. (1992). Shadowing. *Meta: Journal des traducteurs/Meta: Translators' Journal*, *37*(2), 263-273.

Lerner, S. (1975). *A study of shadowing ability*. [Doctoral dissertation, ProQuest Information & Learning).

Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73*(1), 13-22.

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

Lin, C. Y., Wang, M. I. N., Idsardi, W. J., & Xu, Y. I. (2014). Stress processing in Mandarin and Korean second language learners of English. *Bilingualism: Language and Cognition*, *17*(2), 316-346.

Liu, Q. T. (2017). Phonetic Accommodation to Non-Native English Speech. *UC Berkeley PhonLab Annual Report*, *13*(1).

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

Lotz, J., Abramson, A. S., Gerstman, L. J., Ingemann, F., & Nemser, W. J. (1960). The perception of English stops by speakers of English, Spanish. Hungarian, and Thai: A tape-cutting experiment. *Language and speech*, *3*(2), 71-77.

Luo, D., Yamauchi, Y., & Minematsu, N. (2010). Speech analysis for automatic evaluation of shadowing. In *Second Language Studies: Acquisition, Learning, Education and Technology*.

MacLeod, B. (2021). Problems in the Difference-in-Distance measure of phonetic imitation. *Journal of Phonetics*, *87*, 101058.

Manker. (2016). The effect of structural context on phonetic accommodation. *The Journal of the Acoustical Society of America, 139*(4), 2124–2124. https://doi.org/10.1121/1.4950324

Markham, D. (1997). *Phonetic imitation, accent, and the learner* (Vol. 33). Lund University.

Martinsen, R., Montgomery, C., & Willardson, V. (2017). The effectiveness of video-based shadowing and tracking pronunciation exercises for foreign language learners. *Foreign Language Annals*, *50*(4), 661-680.

Mason, I. (2000). Audience design in translating. *The translator*, *6*(1), 1-22.

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models (2nd ed.)*. New York, NY: Chapman and Hall.

Mitleb, F. (1984). Vowel length contrast in Arabic and English: A spectrographic test. *Journal of Phonetics*, *12*(3), 229-235.

Mitterer. (2011). Social accountability influences phonetic alignment. *The Journal of the Acoustical Society of America, 130*(4), 2442–2442. https://doi.org/10.1121/1.3654795

Mori, Y. (2011). Shadowing with oral reading: Effects of combined training on the improvement of Japanese EFL learners‟ prosody. *Language Education & Technology, 48*, 1-22.

Mukherjee, Badino, L., Hilt, P. M., Tomassini, A., Inuggi, A., Fadiga, L., Nguyen, N., & D'Ausilio, A. (2019). The neural oscillatory markers of phonetic convergence during verbal interaction. *Human Brain Mapping, 40*(1), 187–201. https://doi.org/10.1002/hbm.24364

Munro, M. J., & Derwing, T. M. (1995). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and speech*, *38*(3), 289-306.

Murphey, T. (2001). Exploring conversational shadowing. *Language teaching research*, *5*(2), 128-155.

Namy, L. L., Nygaard, L. C., & Sauerteig, D. (2002). Gender differences in vocal accommodation: The role of perception. *Journal of Language and Social Psychology*, *21*(4), 422-432.

Nenkova, A., Gravano, A., & Hirschberg, J. (2008, June 15–20). High frequency word entrainment in spoken dialogue [Paper presentation], (pp. 169-172).. In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, The Ohio State University Columbus, Ohio, USA. (pp. 169-172).

Nguyen, N., & Delvaux, V. (2015). Role of imitation in the emergence of phonological systems. *Journal of Phonetics*, *53*, 46-54.

Nielsen, K. (2011). Specificity and abstractness of VOT imitation. *Journal of Phonetics*, *39*(2), 132-142.

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*, 1496-1509.

Nye, P. W., & Fowler, C. A. (2003). Shadowing latency and imitation: the effect of familiarity with the phonetic patterning of English. *Journal of Phonetics*, *31*(1), 63-79.

Olmstead, A. J., Viswanathan, N., Cowan, T., & Yang, K. (2021). Phonetic adaptation in interlocutors with mismatched language backgrounds: A case for a phonetic synergy account. *Journal of Phonetics*, *87*, Article 101054.

Ostrand, R., & Chodroff, E. (2021). It's alignment all the way down, but not all the way up: Speakers align on some features but not others within a dialogue. *Journal of phonetics*, *88*, Article 101074.

Pardo, J. (2013). Measuring phonetic convergence in speech production. *Frontiers in psychology*, *4*, 559.

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America, 119*, 2382–2393.

Pardo, J. S., Cajori Jay, I., Hoshino, R., Hasbun, S. M., Sowemimo-Coker, C., & Krauss, R. M. (2013). First impressions matter: The influence of role-switching on conversational interaction. *Discourse Processes, 50*, 276-300.

Pardo, J. S., Gibbons, R., Suppes, A., & Krauss, R. M. (2012). Phonetic convergence in college roommates. *Journal of Phonetics*, *40*(1), 190-197.

Pardo, J. S., Urmanche, A., Wilman, S., & Wiener, J. (2017). Phonetic convergence across multiple measures and model talkers. *Attention, Perception, & Psychophysics*, *79*(2), 637-659.

Park, H., & de Jong, K. J. (2008). Perceptual category mapping between English and Korean prevocalic obstruents: Evidence from mapping effects in second language identification skills. *Journal of Phonetics*, *36*(4), 704-723.

Perrachione, T. K., Lee, J., Ha, L. Y., & Wong, P. C. (2011). Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. *The Journal of the Acoustical Society of America*, *130*(1), 461-472.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences, 27*(02), 169-190.

Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency. *Frequency and the emergence of linguistic structure*, *45*, 137.

Pierrehumbert, J. B. (2003). Phonetic diversity, statistical learning, and acquisition of phonology. *Language and speech*, *46*(2-3), 115-154.

Pisoni, D. B., & Lively, S. E. (1995). Variability and invariance in speech perception: A new look at some old problems in perceptual learning. *Speech perception and linguistic experience: Issues in cross-language speech research*, 433-459.

Pouplier, M., Marin, S., & Waltl, S. (2014). Voice onset time in consonant cluster errors: Can phonetic accommodation differentiate cognitive from motor errors?. *Journal of Speech, Language, and Hearing Research*, *57*(5), 1577-1588.

Priva, U. C., & Sanker, C. (2019). Limitations of difference-in-difference for measuring convergence. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *10*(1)

Probst, K., Ke, Y., & Eskenazi, M. (2002). Enhancing foreign language tutors–in search of the golden speaker. *Speech Communication*, *37*(3-4), 161-173.

Purnell, T. C. (2009). The vowel phonology of urban southeastern Wisconsin. *Publication of the American Dialect Society*, *94*(1), 191-217.

Purnell, T., Raimy, E., & Salmons, J. (2017). *Upper Midwestern English*. In R. Hickey (Ed.). *Listening to the Past: Audio records of accents of English* (pp. 298-324). Cambridge: Cambridge University Press.

Qin, Z., Chien, Y., & Tremblay, A. (2016). Processing of word-level stress by Mandarin-speaking second language learners of English. *Applied Psycholinguistics, 38*, 541 - 570.

Raphael, L. J. (2021). Acoustic cues to the perception of segmental phonemes. *The handbook of speech perception*, 603-631.

Reitter, D., & Moore, J. D. (2014). Alignment and task success in spoken dialogue. *Journal of Memory and Language*, *76*, 29-46.

Rojczyk, A. (2012, August 24-25). Phonetic imitation of L2 vowels in a rapid shadowing task [Paper presentation], (pp. 66-76). In Proceedings of the 4th Pronunciation in Second Language Learning and Teaching Conference. Simon Fraser University, Vancouver, British Columbia, Canada**.**

Rost, M., & Wilson, J. J. (2013). *Active listening*. Routledge: New York.

Saadah, E. (2011). *The production of Arabic vowels by English L2 learners and heritage speakers of Arabic***. [**Doctoral dissertation, University of Illinois at Urbana-Champaign]

Saltuklaroglu, T., & Kalinowski, J. (2011). The inhibition of stuttering via the perceptions and production of syllable repetitions. *International Journal of Neuroscience,121*(1), 44-49.

Sancier, M. L., & Fowler, C. A. (1997). Gestural drift in a bilingual speaker of Brazilian Portuguese and English. *Journal of phonetics*, *25*(4), 421-436.

Shepard, C. A., Giles, H., & Le Poire, B. A. (2001). Communication accommodation theory. In W. P. Robinson & H. Giles (Eds.), *The new handbook of language and social psychology* (pp. 33-56). New York: Wiley.

Shockley, Kevin, Laura Sabadini, & Carol A. Fowler. 2004. Imitation in shadowing words. *Perception and Psychophysics 66*, 422– 429.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420

Silva, D. (2006). Acoustic evidence for the emergence of tonal contrast in contemporary Korean. *Phonology, 23*(2), 287-308.

Smith, & Patterson, R. D. (2005). The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age. *The Journal of the Acoustical Society of America*, *118*(5), 3177–3186.

Stoet, G. (2010). PsyToolkit - A software package for programming psychological experiments using Linux. *Behavior Research Methods, 42(4)*, 1096-1104.

Stoet, G. (2017). PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology, 44(1)*, 24-31.

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

Thomson, R., Murachver, T., & Green, J. (2001). Where is the gender in gendered language?. *Psychological Science*, *12*(2), 171-175.

Tobin. (2013). Phonetic accommodation in Spanish-English and Korean-English bilinguals. *The Journal of the Acoustical Society of America, 133*(5), 3340–3340. https://doi.org/10.1121/1.4805637

Ulbrich, C. (2021). Phonetic Accommodation on the Segmental and the Suprasegmental Level of

Speech in Native–Non-Native Collaborative Tasks. *Language and Speech*. https://doi.org/10.1177/00238309211050094

Ullas, Hausfeld, L., Cutler, A., Eisner, F., & Formisano, E. (2020). Neural correlates of phonetic adaptation as induced by lexical and audiovisual context. *Journal of Cognitive Neuroscience, 32*(11), 2145–2158. https://doi.org/10.1162/jocn_a_01608

Wang, X. (2017, December 2-3). The study of shadowing exercise on improving oral English ability for non-English major college students [Paper presentation]. *The World Conference on Management Science and Human Social Development*, Bangkok, Thailand.

Zając, M., & Rojczyk, A. (2014). Imitation of English vowel duration upon exposure to native and non-native speech. *Poznan Studies in Contemporary Linguistics*, *50*(4), 495-514.

Zeller, C. (1997). The investigation of a sound change in progress: /ae/to/e/in Midwestern American English. *Journal of English Linguistics*, *25*(2), 142-155.

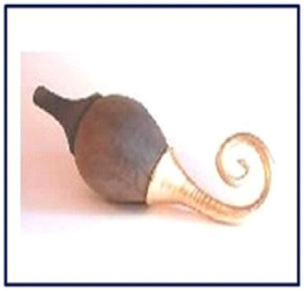
APPENDICES

APPENDIX A: Sentences That Had Real Rhyming Words With Nonsense Words.

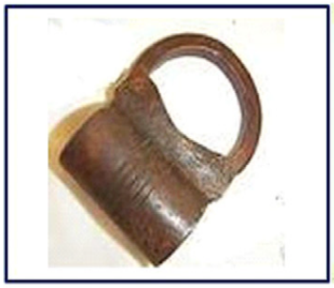
1. duct rhymes with puct
2. dusk rhymes with pusk
3. raffs rhymes with paffs
4. rant rhymes with vant
5. bulb rhymes with vulb
6. wilt rhymes with vilt
7. left rhymes with seft
8. nest rhymes with mest
9. desk rhymes with lesk
10. woke rhymes with roke
11. dote rhymes with sote
12. note rhymes with fote
13. duct rhymes with buct
14. dusk rhymes with busk
15. raffs rhymes with baffs
16. rant rhymes with fant
17. bulb rhymes with fulb
18. wilt rhymes with filt
19. gift rhymes with sift
20. list rhymes with mist
21. disk rhymes with lisk
22. duke rhymes with ruke
23. root rhymes with suit
24. shoot rhymes with fute

APPENDIX B: Images Of Nonobjects and Their Corresponding Nonsense Words.

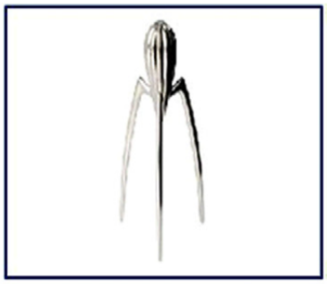
paffs



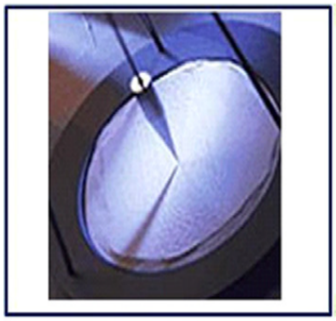
puct



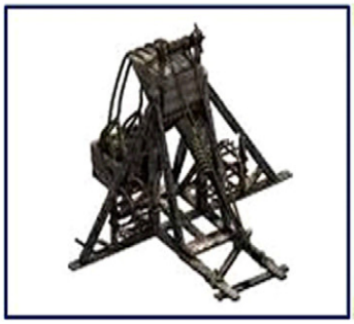
pusk



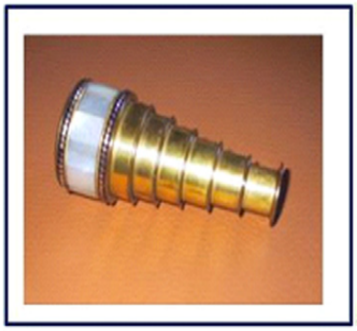
vant



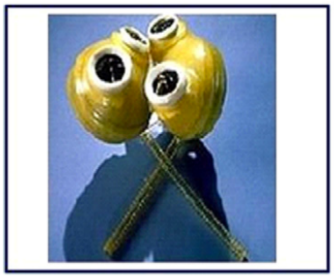
vilt



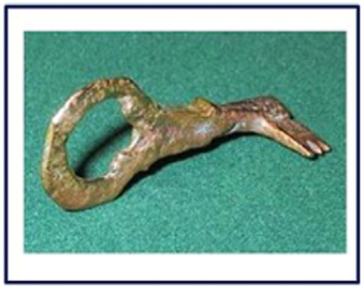
vulb



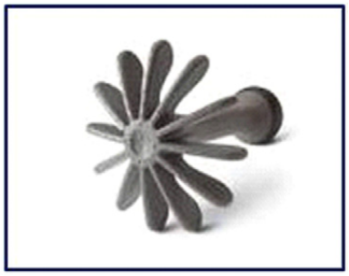
ceft



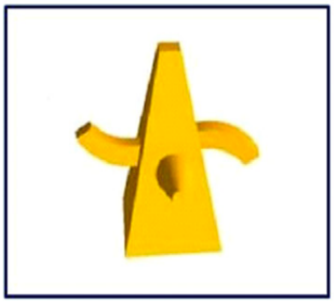
lesk



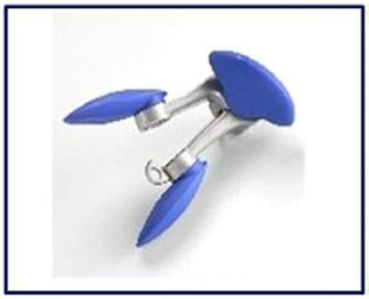
mest



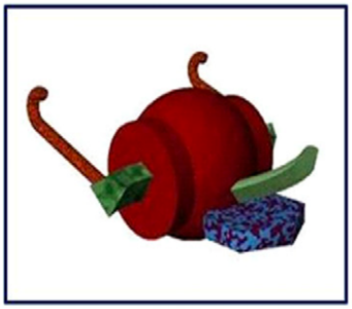
roke



fote



sote



(Bürki, Spinelli, & Gaskell, 2012)

APPENDIX C: Overall Trainees' Intelligibility Improvement Scores in C-Group at the Pre- and Post-Test.

Trainee	All segments		
	Pre-test	Post-test	Difference
3	0.29	0.56	0.27
6	0.43	0.44	0.01
11	0.58	0.79	0.21
13	0.26	0.6	0.34
14	0.09	0.64	0.55
15	0.54	0.67	0.13
21	0.53	0.89	0.36
24	0.68	0.77	0.10
25	0.45	0.79	0.34
26	0.29	0.68	0.39

Note. these data were averaged across the four trained segments.

APPENDIX D: Overall Trainees' Intelligibility Improvement Scores in D-Group at the Pre- and Post-Test.

Trainee	All segments		
	Pre-test	Post-test	Difference
1	0.45	0.68	0.23
2	0.40	0.47	0.07
8	0.46	0.87	0.41
9	0.56	0.56	0
16	0.60	0.68	0.08
18	0.70	0.83	0.13
20	0.56	0.69	0.13
22	0.36	0.70	0.34
23	0.53	0.73	0.20
27	0.51	0.68	0.17

Note. these data were averaged across the four trained segments.

APPENDIX E: Trainees' Intelligibility Improvement Scores of /p/ in C-group at the Pre- and Post-Test.

Trainee	/p/		
	Pre-test	Post-test	Difference
3	.25	.45	.20
6	.27	.13	-.13
11	.98	.93	-.05
13	.08	.45	.37
14	.05	.38	.33
15	.42	.48	.07
21	.93	.98	.05
24	.42	.93	.52
25	.48	.75	.27
26	.00	.67	.67

APPENDIX F: Trainees' Intelligibility Improvement Scores of /p/ in D-Group at the Pre- and Post-Test.

Trainee	/p/		
	Pre-test	Post-test	Difference
1	.12	.58	.47
2	.47	.33	-.13
8	.08	1.00	.92
9	.90	.97	.07
16	.33	.53	.20
18	1.00	1.00	.00
20	.33	.65	.32
22	.13	.62	.48
23	.85	.72	-.13
27	.65	.75	.10

APPENDIX G: Trainees' Intelligibility Improvement Scores of /v/ in C-Group at the Pre- and Post-Test.

Trainee	/v/		
	Pre-test	Post-test	Difference
3	.65	.88	.23
6	.93	.93	.00
11	.90	.87	-.03
13	.25	.82	.57
14	.03	.67	.63
15	.80	.95	.15
21	.85	.82	-.03
24	1.00	.95	-.05
25	.98	.98	.00
26	1.00	.98	-.02

APPENDIX H: Trainees' Intelligibility Improvement Scores of /v/ in D-group at the Pre- and Post-Test.

Trainee	/v/		
	Pre-test	Post-test	Difference
1	.93	.92	-.02
2	.87	.80	-.07
8	.93	.93	.00
9	.92	.65	-.27
16	.93	.92	-.02
18	.92	.93	.02
20	.83	.85	.02
22	.97	.93	-.03
23	.97	.98	.02
27	.70	.45	-.25

APPENDIX I: Trainees' Intelligibility Improvement Scores of /ɛ/ in C-group at the Pre- and Post-Test.

Trainee	/ɛ/		
	Pre-test	Post-test	Difference
3	.00	.28	.28
6	.02	.10	.08
11	.03	.60	.57
13	.40	.52	.12
14	.00	.80	.80
15	.32	.53	.22
21	.13	.97	.83
24	.50	.42	-.08
25	.05	.72	.67
26	.02	.47	.45

APPENDIX J: Trainees' Intelligibility Improvement Scores of /ɛ/ in D-group at the Pre- and Post-Test

Trainee	/ɛ/		
	Pre-test	Post-test	Difference
1	.38	.78	.40
2	.02	.08	.07
8	.38	.85	.47
9	.13	.17	.03
16	.87	.82	-.05
18	.60	.68	.08
20	.80	.92	.12
22	.03	.47	.43
23	.15	.55	.40
27	.37	.88	.52

APPENDIX K: Trainees' Intelligibility Improvement Scores of /oʊ/ in C-group at the Pre- and Post-Test.

Trainee	/oʊ/		
	Pre-test	Post-test	Difference
3	.27	.63	.37
6	.52	.60	.08
11	.40	.75	.35
13	.30	.63	.33
14	.27	.72	.45
15	.63	.70	.07
21	.22	.80	.58
24	.78	.78	.00
25	.27	.70	.43
26	.13	.60	.47

APPENDIX L: Trainees' Intelligibility Improvement Scores of /oʊ/ in D-group at the Pre- and Post-Test

Trainee	/oʊ/		
	Pre-test	Post-test	Difference
1	.35	.42	.07
2	.23	.65	.42
8	.43	.70	.27
9	.28	.45	.17
16	.28	.47	.18
18	.30	.70	.40
20	.27	.35	.08
22	.30	.78	.48
23	.15	.68	.53
27	.33	.63	.30

CURRICULUM VITAE

Ruqayyah Althubyani

Place of Birth: Taif, Saudi Arabia

Education

PhD in Linguistics

University of Wisconsin-Milwaukee (2021)

MA in Linguistics

University of Florida Atlantic University (2015)

Certificate in the teaching of English as a second language (TESOL)

University of Florida Atlantic University (2015)

BA in English Language and Literature

College of Education, Taif (2006)

Experience

Umm Al-Qura University

Teaching Assistant in the Department of English (2008- 2011)

Saudi Public School

English Teacher (2006 – 2007)

British School (Taif, Saudi Arabia)

English Teacher (2007 – 2008)

Conferences

Althubyani, R. & Park, H. (2019, October 4-5). The Effect of Dialects on Phonetic Convergence in Non-native Settings [Poster presentation]. The meeting of the 24th Annual Mid-Continental Phonetics & Phonology Conference, University of Wisconsin-Milwaukee, Milwaukee, WI, United States.