WEIGHING PHONETIC PATTERNS IN NON-NATIVE ENGLISH SPEECH

by

Zhiyan Gao
A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
In Partial fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Linguistics

Committee:

_____ Director

_____

_____

_____

_____ Department Chairperson

_____ Program Director

_____ Dean, College of Humanities and Social Sciences

Date: _____ Fall Semester 2019
George Mason University
Fairfax, VA

Weighing Phonetic Patterns in Non-Native English Speech

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

By

Zhiyan Gao
Master of Arts
George Mason University, 2012
Bachelor of Arts
Soochow University (Suzhou, China), 2009

Director: Dr. Steven Weinberger, Associate Professor
Department of English

Fall Semester 2019
George Mason University
Fairfax, VA

Acknowledgments

I would like to express my deepest gratitude and appreciation to my dissertation committee who allowed an invaluable insight into their knowledge. I am also grateful to George Mason University, which funded this dissertation via a Presidential Scholarship and three summer research grants.

I am deeply indebted to my advisor Dr. Steven Weinberger who introduced me to the field of foreign accent and guided me every step of the way as I formed the ideas for the dissertation, designed experiments, collected data, and started writing. I also thank Dr. Weinberger for providing me with an assistantship and full access to the Speech Accent Archive. This dissertation would not have been possible without Dr. Weinberger.

I am extremely grateful to Dr. Douglas Wulf for his insightful comments and valuable feedback throughout the development of this dissertation. I would also like to extend my gratitude to Dr. Harim Kwon for helping me understand theories of speech perception and for her advice on acoustic analysis and statistical reasoning. I am grateful to Dr. Dennis Perzanowski for his encouragement and guidance during my early years in graduate school. I greatly appreciate Dr. Perzanowski's help throughout the years and his invaluable comments on various aspects of this dissertation.

I sincerely thank Dr. Jennifer Culbertson for introducing me to R and Amazon Mechanical Turk, both of which were utilized in this dissertation. I also would like to thank Dr. Tuuli Morrill for taking me as her research assistant. I thank Dr. Morrill for teaching me various experimental techniques and allowing me to try out my ideas using her research money.

I would like to thank Dr. Rias van den Doel (Utrecht University) for his insightful comments on an early draft of this dissertation. I thank my friend Ray Varner for his comments on an early version of the demographic questionnaire designed for this project. I also thank my friend David Blau for providing an outsider's opinion on this project.

I am grateful to my fellow classmates Dr. Abdullah Alfaifi, Mashael Al-Aloula, Omar Alkhonini, Yuting Guo, and Chiu-ching Tseng for their help and friendship. I also thank my former flatmate Dr. Jitendra Parajuli for sharing his books with me and introducing me to the world of philosophy and rock and roll.

I thank my mother Minli Gu, my father Lijun Gao, my grandmother Shuyuan Chen, and my late grandfather Wenfeng Gao for their unconditional love and support. This dissertation is dedicated to them.

Table of Contents

List of Tables

List of Figures

Abstract

WEIGHING PHONETIC PATTERNS IN NON-NATIVE ENGLISH SPEECH

Zhiyan Gao, PhD

George Mason University, 2019

Dissertation Director: Dr. Steven Weinberger

Non-native (L2) speakers of English often speak English with a certain degree of foreign accent. While much research has investigated the accentedness of L2 English speech, very few studies have attempted to rank phonetic patterns in L2 speech according to their perceived foreign accentedness. This dissertation provides accentedness rankings of a large variety of phonetic patterns in L2 English speech. By investigating why some phonetic patterns in L2 English speech are perceptually more accented than others, this dissertation reveals the possible underlying mechanisms that govern foreign accent perception.

This dissertation focuses specifically on the segmental and syllable structural aspects of L2 speech. Two perception experiments were conducted to elicit native (L1) American English raters' accentedness judgments on 100 L2 stimuli extracted from the Speech Accent Archive (Weinberger, 2019). In both experiments, raters heard L2 English speech samples and rated foreign accentedness for each speech sample. Accentedness rankings of various phonetic patterns in L2 speech were therefore obtained. Linear mixed-effects models were implemented to investigate the effect of different types of phonetic patterns on accentedness perception. Prosodic information of the stimuli was accounted for in the least intrusive manner by employing a Dynamic Time Warping approach. Results of the two experiments show that (1) the consonant and syllable structural aspects of L2

speech carry more weight in accentedness perception than vowels; (2) phonological context affects accentedness perception; (3) raters are aware of which phonetic patterns are allowed in L1 speech, showing that L1 phonetic and phonological knowledge (L1 knowledge) might affect accentedness perception.

A third experiment was further conducted to investigate how raters' L1 knowledge affected their accentedness judgment. A Naïve Discriminative Learning Model (NDL) was employed to approximate raters' L1 knowledge by examining the co-occurrences of pronunciations (e.g., [æsk]) and lexical outcomes (e.g., "*ask*") in L1 speech of American English. 100 L2 stimuli were subsequently evaluated against listeners' L1 knowledge to estimate the degree of dissimilarity (NDL-distance) between L1 and L2 speech samples. The results show that NDL-distance correlates significantly with accentedness ratings, suggesting that L1 knowledge, as approximated by the NDL model, could have affected accentedness judgments.

This dissertation contributes to the field of foreign accent by providing accentedness rankings of various phonetic patterns in L2 speech. In lieu of ad hoc explanations for why some phonetic patterns are more accented than others, this dissertation directly examines how raters' L1 knowledge affected their accentedness judgments of L2 speech, providing insights into the nature of foreign accent perception.

Chapter 1: Introduction

"Foreign accent" is usually considered an issue of perception, rather than production. Only perceivable deviations in non-native (L2) speech are considered features of "foreign accent." As Munro and Derwing (1998, p.160) defines it, foreign accent is "*the extent to which an L2 learner's speech is perceived to differ from native speaker norms.*" Foreign accent has been a widely discussed issue in the literature, and consonant manner of articulation (e.g., Riney and Takagi, 1999; Solon, 2015), vowel quality (e.g., Hahn, 2004; Munro and Derwing, 2001; Zielinski, 2008) and various prosodic elements (e.g., Hahn, 2004; Munro and Derwing, 2001; Zielinski, 2008) have been found to greatly affect foreign accent perception. However, there seems to be no consensus on which elements contribute more to foreign accentedness. The current study aims to explore the degree of foreign accentedness exhibited by various phonetic/phonological patterns in L2 speech. Specifically, we investigate which phonetic/phonological aspects of L2 English speech are perceived as more saliently foreign-sounding by native (L1) listeners of American English. The current study also endeavors to uncover possible reasons for why some phonetic and phonological patterns in L2 speech are perceived as more foreign-sounding than others. The focus of the current study is L1 and L2 English speech. The term "accent" used in this dissertation refers specifically to L1 and/or L2 English speech patterns. The cited studies, unless otherwise noted, are research on L1 and/or L2 English speech.

The current study investigates both phonetic and phonological differences between L1 and L2 English speech. For example, using [æ̟] instead of [æ] could be an issue of phonetics, while structural changes such as vowel epenthesis (e.g., [æskə] for "*ask*") could be an issue of phonology. However, we consider all these differences an issue of phonetics, because we assume that the L2 speakers who produced these stimuli were aware of and were trying to mimic L1 productions. In other words, we assume that the L2 speakers know how the phonetic segments should be organized in English (i.e. phonology). We consider the differences between the L2 speech samples and their corresponding

L1 productions as a reflection of the differences between L1 and L2 speakers' articulatory configurations, which should be discussed in terms of phonetics. Therefore, the term "phonetic pattern" is used throughout the dissertation to refer to pronunciation differences between L1 and L2 speech. We fully acknowledge that some of the phonetic patterns discussed in this dissertation could be considered issues of phonology.

1.1 Background

According to Wells (1982), a difference between varieties of English may involve any or all aspects of the language (e.g., syntax, pronunciation, and lexicon etc.), while a difference between accents of English is restricted to pronunciation. Since the current study aims to investigate the accentedness of various speech patterns, the focus is on the phonetic and phonological aspects of L1 and L2 utterances.

It should be noted that an accent is not unique to L2 speakers. It is widely accepted among linguists that an accent is something both L1 and L2 speakers have (Lippi-Green, 2012). However, L1 and L2 accents could lead to different sociolinguistic consequences. People with an L1 accent are often perceived by other L1 speakers as being more trustworthy and capable than people with an L2 accent (Gluszek and Dovidio, 2010). L2 accents, although bearing no relationship to one's intelligence or personal character, are sometimes viewed as characteristics of ineptitude (Gluszek and Dovidio, 2010). L2 speakers, therefore, place great importance on the accuracy of their pronunciations (Waniek-Klimczak, Rojczyk, and Porzuczek, 2015).

Other than the differences in sociolinguistic consequences, L1 and L2 English accents also exhibit different types of phonetic/phonological patterns (Lippi-Green, 2012). Take t/d-deletion in English as an example. L1 English speakers are more likely to delete /t/ or /d/ when they are past tense morphemes (e.g., /d/ in "called," /t/ in "packed") than when they are part of the stem of a word (e.g., /d/ in "hold," /t/ in "pact") (Guy, 1991). L2 speakers' t/d-deletion strategy, on the other hand, does not seem to concern whether the /t/ or /d/ is part of the stem of a word (Edwards, 2011; Hansen, 2004).

The differences between L1 and L2 English accents could also be discussed from a perceptual

perspective. Previous research found that children as young as seven years old are capable of distinguishing L1 accents from L2 ones (Floccia et al., 2009). They are, however, not as successful at distinguishing other regional L1 accents from their own (Floccia et al., 2009). The reason for such asymmetry was attributed to the type of phonetic variations in L1 and L2 speech. As Floccia et al. (2009) found, L2 speech exhibits greater distortions of consonants than L1 speech. Children in Floccia et al. (2009) seemed to have relied more on consonant information when distinguishing between L1 and L2 accents. This suggests that foreign-accented speech exhibits characteristics that distinguish itself from L1 speech varieties and these characteristics are perceivable.

In the United States, General American English (GA) is often colloquially referred to as the "neutral American accent" or the "standard American accent" (Lippi-Green, 2012; Wells, 1982). However, GA is not a uniform accent (Wells, 1982). According to Wells (1982), the rime of the word "square" in GA could have at least three L1 variations, namely [ɛɹ], [æɹ] and [eɪɹ]; for words such as "star," the vowel could range from a fronted [a] to a back [ɑ]. In other words, the GA is not one specific accent of American English, but a continuum of accents. Although dictionaries and English textbooks tend to teach the "Standard American English," no such standardized English accent has existed in reality (Lippi-Green, 2012).

The heterogeneity among L1 English accents poses a great challenge to empirical inquiries on the differences between L1 and L2 accents. When discussing the hypothetical "native speaker norms" or a "typical native accent," previous research tends to rely on mean acoustic measurements of L1 speech produced by a specific group of L1 speakers (e.g., Chan, Hall, and Assgari, 2016; McCullough, 2013). For example, mean L1 voice onset time (VOT) was used to represent the native speaker norm of aspiration length for plosive consonants (Riney and Takagi, 1999). Mean formant frequencies were used to represent L1 vowel qualities (Chan, Hall, and Assgari, 2016).

With the assumption that mean acoustic measurements represent the native speaker norms, previous perception research has identified several phonetic correlates of foreign accent (e.g., Chan, Hall, and Assgari, 2016; Riney and Takagi, 1999; Solon, 2015). Consonant manner of articulation, VOT, and vowel epenthesis have often been found to be closely associated with perceived foreign accent (Chan, Hall, and Assgari, 2016; Magen, 1998; Munro and Derwing, 1998; Riney and Flege,

1998; Solon, 2015). Several studies also suggested that vowel quality contributes to foreign accent perception (Braun, Lemhöfer, and Mani, 2011; Major, 1986; Park, 2013). Various prosodic elements such as intonation and speech rate have also been found to greatly affect foreign accent perception (Hahn, 2004; Kang, Rubin, and Pickering, 2010; Munro and Derwing, 2001; Zielinski, 2008).

There is no doubt that the degree of perceived foreign accent is affected by various phonetic patterns. However, it is difficult to determine a hierarchy of their importance and it is equally challenging to quantify their relative impact on accentedness detection (Munro and Derwing, 1995; Rognoni and Busà, 2014). Among the few studies that did evaluate the relative impact of different phonetic patterns, the findings were often inconsistent and sometimes contradicted one another (Magen, 1998; van den Doel, 2006). Some of the problems in these studies could be due to experimental design rather than strictly linguistic factors. For example, the stimuli of several previous studies were acoustically edited or synthesized (e.g., Chan, Hall, and Assgari, 2016; Jilka, 2000; Magen, 1998. It is possible that acoustic manipulations in one dimension (e.g., intonation) could have affected the perception of acoustic signals in other dimensions (e.g., vowel height) (Whalen and Levitt, 1995). Therefore, an acoustically edited or synthesized speech sample might not be representative of natural speech.

Some research placed L2 stimuli in carrier sentences without controlling for the effect of phonological context (Magen, 1998; van den Doel, 2006). It is possible that a given L2 stimulus could be perceptually accented in one context but not accented in another, which could potentially explain conflicting results found by previous studies. Further, regarding the training of listeners specifically for the task of judging accents, some studies claim that a training session is necessary to familiarize listeners with the task (e.g., Major, 1986), while others argue that a training session could introduce biases that affect listeners' judgments (e.g., McDermott, 1986). However, there seems to be a paucity of literature that directly compares accentedness judgments between trained and untrained listeners. The current study addresses this methodological gap by using unaltered stimuli and conducting two separate experiments which directly compared trained and untrained listeners.

While some phonetic patterns in L2 speech were found to be perceptually more foreign accented than others, the reason for such a phenomenon is not readily clear (e.g., Magen, 1998). Previous

research has often made ad hoc claims that listeners' knowledge of English (e.g., English phonotactics) is one of the underlying mechanisms that affects accentedness judgments on L2 speech (e.g., Park, 2013). The current study investigates this claim directly by approximating L1 English listeners' knowledge of English phonetics and phonology (L1 knowledge) with a computational model. L2 speech samples were evaluated by the model to generate dissimilarity scores that approximated the degree of difference between L1 and L2 speech samples. Accentedness ratings on the L2 speech samples were compared against the dissimilarity scores to observe how exactly L1 knowledge affects accentedness perception.

1.2 The Current Study

The current study focuses on which phonetic patterns in L2 speech are perceptually more accented to L1 listeners of American English, rather than how acoustic cues correlate with perceived foreign accentedness. For example, the current study investigates whether an L2 pronounciation of "*thick*" as [θik] is perceptually more foreign-accented to an L1 listener than pronouncing "*thick*" as [θɪk]. The degree of acoustic differences between an L2 [i] and its L1 target [ɪ] and their effect on accentedness perception are not a focus of the current study.

In lieu of acoustic differences between L1 and L2 productions, the current study opts to focus directly on perception data. To this end, phonetically trained personnel were recruited to phonetically transcribe L1 and L2 utterances of English using the International Phonetic Alphabet (IPA). IPA symbols represent transcribers' perceptions. Once a sound is transcribed with an aspiration symbol [ʰ] (e.g., [pʰ]), it is safe to conclude that the transcriber perceived it. Furthermore, when several transcribers perceive the same phonetic phenomenon, the perception is corroborated.

The current study conducted two perception experiments to elicit accentedness judgments from L1 listeners of American English. One hundred L2 audio speech samples were selected from the Speech Accent Archive (SAA: Weinberger, 2019) as stimuli for the two experiments. The stimuli were selected based on their IPA transcriptions in the SAA. In classifying the 100 L2 audio stimuli, the current study surveyed productions of 100 L1 speakers of American English. The most common productions among the 100 L1 speakers were considered L1 target productions. For example, 90%

of the L1 productions for the word "*thick*" were rendered as [θɪk]. [θɪk] was therefore selected as the L1 target production for "*thick*." Such a treatment was based on the assumption that L1 English listeners should all be familiar with the most common L1 productions and consequently consider them as containing no foreign accent. L2 speech samples that were transcribed the same as their L1 target production (e.g. [θɪk] for "*thick*") were termed "match" stimuli, meaning that the L2 productions matched their L1 target productions. L2 productions that did not match their L1 target productions were termed "mismatch" stimuli. Both match and mismatch stimuli were produced by L2 English speakers.

The current study focuses specifically on segmental and syllable structural aspects of L2 English speech. Therefore, the mismatches were defined in terms of segmental and syllable structural characteristics. Prosodic elements such as intonation and speech rhythm were ignored in the process of stimuli selection. Prosody of the stimuli was controlled for by selecting stimuli without lexical stress shifts. Intonational and durational information was controlled for computationaly with an alignment algorithm, which estimates intonational and durational differences between L1 and L2 speech samples. Section 4.5 in Chapter 4 discusses this method in more detail.

It should be noted that some of the mismatch productions, although they were not transcribed the same as their L1 target productions, are not unique to L2 speakers of English. For example, two native speakers of American English in the SAA pronounced "*thick*" as [t̪ɪk], which does not match the most common L1 production [θɪk]. [t̪ɪk] was termed by the current study as a mismatch stimulus, simply because it does not match the most common L1 production. Presumably, mismatch stimuli with phonetic patterns that exist in L1 speech would be judged as less accented. On the other hand, some mismatch stimuli contain phonetic patterns that do not occur in L1 speech (e.g., pronouncing "*ask*" as [æskə]). These stimuli presumably carry a relatively higher degree of accentedness.

The mismatch stimuli were further divided into three groups based on three types of mismatches, namely, stimuli with consonant mismatches (e.g., [tɪk] for "*thick*"), stimuli with vowel mismatches (e.g., [θik] for "*thick*"), and stimuli with syllable structure mismatches (e.g., [æskə] for "*ask*"). The current study therefore investigated four types of stimuli (i.e., the match stimuli and the three types of mismatch stimuli). L1 listeners of American English (i.e., the raters) were recruited to rate the

foreign ccentedness of the four types of stimuli. Results show that the mismatch stimuli were judged as being more accented than the match stimuli. Among the three types of mismatch stimuli, stimuli with consonant and syllable mismatches were judged as being the most accented, and stimuli with vowel mismatches were judged as the least accented. In addition, the frequency of occurrences of a phonetic pattern in L1 speech potentially affects accentedness judgment. Chapter 4, Chapter 5 and Chapter 6 of this dissertation discuss the findings in more detail.

To investigate how raters' L1 phonetic and phonological knowledge (L1 knowledge) affects their accentedness judgment, the current study computationally constructed an L1 production model based on IPA transcriptions from 100 L1 speakers of American English. The L1 production model was a matrix of "association strengths," approximating the probable occurrence of a certain word based on a certain sequence of phonetic segments. For example, the model approximated the probability of the segment sequence [p^{h}li] occurring as a prediction of the word "*please*." The model could be intuitively understood as the probability that a rater believed that the hearing of the [p^{h}li] portion of the word resulted in the intended meaning of "*please*." The L1 production was therefore a matrix of "association strengths," mapping phonetic segment sequences to lexical meanings.

IPA transcriptions of the 100 L2 stimuli were compared against the L1 production model to generate dissimilarity scores, representing the degree of difference between an L2 stimulus and the 100 L1 productions. Dissimilarity scores and acccentedness ratings of the L2 stimuli were compared against each other to evaluate whether the dissimilarity scores could predict the degree of foreign accentedness. The results show that the dissimilarity scores indeed correlate significantly with accentedness ratings, indicating that L1 knowledge, as approximated by the L1 production model, could have significantly affected accentedness judgments. The L1 production model and the subsequent comparisons between L1 and L2 speech samples provide insights into the nature of L1 knowledge and how it affects accentedness perception. Chapter 6 of this dissertation provides further details.

1.3 Rationale

The current study hypothesizes that some phonetic patterns in L2 speech are perceptually more accented than others. Specifically, stimuli with consonant mismatches are perceptually more accented than stimuli with vowel mismatches. This hypothesis is grounded in previous studies on lexical identification and speech perception (e.g., Kronrod, Coppess, and Feldman, 2012; Nespor, Peña, and Mehler, 2003). These studies often claim that consonants are more important in lexical identification and are generally more susceptible to categorical perception. These two potential attributes of consonants could have affected the accentedness of stimuli with consonant mismatches.

Speech processing involves the ability to detect statistical regularities in the input (Romberg and Saffran, 2010). Both adults and children have been shown to be able to parse speech by extracting statistical regularities such as transitional probabilities between phonetic segments (Romberg and Saffran, 2010). However, not all statistical regularities are equally weighted in every circumstance. Nespor, Peña, and Mehler (2003) suggest that there is a division of labor between consonants and vowels in language processing and acquisition. Transitional probabilities between consonants are more important at the lexical level, while transitional probabilities of vowels are more important at prosodic and syntactic levels (Nespor, Peña, and Mehler, 2003). The distortion of consonants would therefore be more likely to impair lexical identification, while the distortion of vowels would affect the processing of prosody and syntax. If lexical identification is a component of perceived foreign accentedness, then consonant mismatches, especially the ones that do not represent L1 dialectal variations, would be perceived as more accented than vowel mismatches. Alternatively, if prosodic and syntactic information weighs more heavily than lexical identification in accentedness judgment, then vowel mismatches would perhaps be more accented than consonant mismatches.

Consonants and vowels not only differ in their potential functions in lexical identification, but also in how they are perceived. Research on speech perception showed that phonemic differences between consonants are more easily perceived as a categorical difference than phonemic differences between vowels (Altmann et al., 2014; Kronrod, Coppess, and Feldman, 2012). Perception of consonants, especially obstruent consonants, is relatively more categorical, while the perception of vowels

is relatively more continuous (Altmann et al., 2014; Kronrod, Coppess, and Feldman, 2012). Therefore, phonemic alternations of consonants are probably more accented, because such alternations are more easily perceived as a categorical difference.

The difference between consonants and vowels could also be discussed from phonological-distributional perspectives. Unlike consonants, vowels are cross-linguistically fewer in number and more prone to lose their distinctiveness (Nespor, Peña, and Mehler, 2003). If vowel variations are indeed more common than consonant variations in L1 speech, then vowel variations in L2 speech could be more tolerable than consonant variations. Consequently, stimuli with vowel mismatches could be perceived as less accented than stimuli with consonant mismatches.

Studies and theories on lexical identification and speech sound categorization do not provide much evidence for the accentedness of stimuli with syllable mismatches, such as segment deletion and segment epenthesis. However, since segment epenthesis is less common than segment deletion in L1 English speech (Johnson, 2004b), it is possible that stimuli with segment deletion is less accented than stimuli with segment epenthesis.

1.4 Organization of the Dissertation

This dissertation proceeds as follows: Chapter 2 begins with a literature review on certain phonetic/phonological aspects of foreign accents, theories of language perception, and experimental methodologies. Chapter 2 summarizes major findings in previous studies and discuss their advantages and shortcomings. Chapter 3 discusses how the stimuli for the experiments were selected.

Chapter 4 discusses Experiment 1, which was conducted as a pilot study. Major findings and potential methodological shortcomings are discussed. Accentedness rankings of phonetic patterns in L2 speech are provided. Discussions in this chapter center on the potential reasons for the observed accentedness ratings. Chapter 5 considers Experiment 2, which addresses the potential methodological problems in Experiment 1. In addition to major findings of the experiment, discussions in this chapter considers how the methodological differences between Experiment 1 and Experiment 2 affected accentedness judgment.

Chapter 6 discusses Experiment 3. The aim of Experiment 3 is to investigate whether raters' L1

knowledge affects their accentedness judgment. A computational model was adopted to construct an L1 production model, which approximated the degree of phonetic and phonological differences between L1 and L2 speakers' speech samples. The dissimilarity scores generated by the model were compared against accentedness ratings from Experiment 2 to investigate how L1 knowledge affects accentedness judgment. This chapter further discusses the successes and failures of the model in approximating accentedness judgment.

Chapter 7 considers implications of the findings, problems of the current study, and offers suggestions for future research.

Chapter 2: Literature Review

The aim of the current study is to investigate which phonetic patterns in non-native (L2) English speech are judged to be foreign accented by native (L1) listeners of American English. This chapter provides an overview of research on phonetic and phonological characteristics of foreign accent, after which the focus shifts to relevant theories of speech perception and lexical identification. Based on the findings and theories in previous literature, the rationale for the hypothesis of the current study is further discussed.

2.1 Segmental Correlates of Foreign Accent

Previous literature often suggests that consonant variations are associated with perceived accentedness (e.g., Flege and Munro 1994; Magen 1998). Major (1987) found that L2 English speakers' production of English Voice Onset Time (VOT) correlates with perceived accentedness. He found that the closer the L2 VOT conforms to the American English norm (i.e., mean L1 VOT), the higher the accentedness score (i.e., the more native-like a speaker is perceived). González-Bueno (1997) similarly showed that VOT affects the degree of perceived accentedness. In her study, the Spanish /k/s produced by English speakers with a VOT outside of an L1 Spanish speaker's range, which is between 15 to 35 milliseconds (ms), were rated as very accented by monolingual Spanish speakers. This suggests that L1 Spanish speakers are also sensitive to the temporal boundaries of plosive consonants.

L2 liquid consonant variations (e.g., using [ɾ, ɾ] instead of [l, ɹ]) might also be associated with foreign accent. For example, the substitution of the Japanese flap (i.e., [ɾ]) for the English liquids [l] and [ɹ] was found to be indicative of foreign accent by L1 English speakers (Riney, Takada, and Ota, 2000). Solon (2015) studied English speakers' L2 pronunciation of the Spanish [l]. When English speakers pronounced a Spanish [l] as a velarized English [lˠ], the sound was rated as very foreign

accented by L1 Spanish speakers. Solon (2015) further showed that the degree of foreign accentedness of the non-Spanish-like [lˠ] depends on the degree of velarization, as measured by the frequency of the second formant (F2) and the duration of the L2 [lˠ]. The lower the F2, the shorter the non-native [lˠ] and the more foreign-accented the L2 [lˠ] sounds to native Spanish speakers. Although previous research did show that consonant variations (e.g., VOT-shortening, [ɹ, l] to [ɾ]) could affect accentedness judgments, the types of consonant variations discussed in previous literature were limited, with most of the research focusing on plosives and liquids. To further validate the effect of consonant variations on foreign accent perception, evidence is needed for whether other types of consonant variations (e.g., feature changes in fricatives) lead to similar accentedness judgments.

Major (1987) focused on vowel perceptions. He showed that foreign accentedness might be associated with some vowels, but not with others. In his study, L1 American English listeners first provided accentedness ratings of L2 productions of the English word "*bet*" ([bɛt]) and "*sat*" ([sæt]). The initial consonants were then edited out, leaving only the vowels and the codas (i.e., [ɛt] and [æt]). Another group of L1 English listeners were presented with these stimuli (i.e., [ɛt]s and [æt]s) in an identification test. The task was to determine whether the stimulus they heard was [ɛt] or [æt]. Accuracy ratings were calculated based on how often a stimulus was correctly identified (e.g., how often an L2 [ɛt] was identified as [ɛt] rather than [æt]).

The study showed that the accentedness ratings of [sæt] negatively correlated with the accuracy ratings of its rime (e.g., [æt]). That is, the more accurately the L2 [æt] was identified by L1 English listeners as [æt], the less accented [sæt] sounded. However, the accentedness of [bɛt] positively correlated with the accuracy of its rime [ɛt], suggesting that [bɛt] was rated as being more accented when its rime [ɛt] was more accurately identified as [ɛt]. Major (1987) thus concluded that whether the pronunciations of vowels have any impact on perceived accentedness depends on the specific vowel in question.

Similar findings were reported by Chan, Hall, and Assgari (2016), who used synthesized English phonemes /a/, /æ/ and /ʌ/ as stimuli. In their study, the frequencies of the first (F1) and second formants (F2) and durations of the three vowels were manipulated. L1 English speakers were recruited to provide accentedness judgments on the synthesized stimuli. The results show that the deviations

of F1 and F2 from mean L1 frequency values might increase the perceived accentedness for some vowels but decrease the perceived accentedness for other vowels. For example, as F1 and F2 center frequency values became lower than L1 formant frequency values (i.e., mean center frequencies), /a/ and /æ/ were rated as more accented. Surprisingly, when F1 and F2 values of /ʌ/ became lower than mean center formant frequencies of L1 speakers, /ʌ/ became less accented. In other words, the deviation of vowel formant frequencies may either increase or decrease accentedness depending on the specific vowel in question.

The reason for this phenomenon was attributed to the overlap of vowel space. As the F1 and F2 of /a/ and /æ/ became lower with respect to L1 formant frequency values, /a/ and /æ/ started to overlap with L1 formant frequencies of /ʌ/ and /a/, respectively. In other words, the lowering of F1 and F2 for /a/ and /æ/ made them sound like other vowels to L1 speakers. However, when F1 and F2 values of /ʌ/ became lower, /ʌ/ actually moved away from L1 formant frequencies of /a/ and /æ/. As a result, the degree of overlapping decreased, which in turn decreased the perceptual accentedness. Contrary to findings in Chan, Hall, and Assgari (2016) and Major (1987), McCullough (2013) found that greater formant deviation from L1 speaker norms leads to higher degrees of accentedness, no matter which vowel is considered. This finding is consistent with several other studies, which also showed independent effects of both static F1 and dynamic/static F2 values on perceived accentedness (Munro, 1993; Wayland, 1997).

Previous research on L2 vowel perception has shown conflicting findings. It could be the case that the L1 speaker norm, from which vowel deviation is usually measured, is not a representative indicator (Chan, Hall, and Assgari, 2016). Perhaps vowel perception depends on the spectral overlapping of vowel categories. Since there is no agreement on which acoustic characteristics are responsible for perceived "vowel errors," problems might emerge when one tries to synthesize "vowel errors."

To avoid the complexity of the relationship between acoustic signals and phonemic categories, the current study opts to focus directly on perception data, rather than acoustic signals in L2 speech. Phonetically trained personnel were recruited to transcribe L2 English speech using the International Phonetic Alphabet (IPA). The IPA transcriptions represent transcribers' perceptions. For example,

once an L2 production of "*thick*" is transcribed as [θik], one can assume that the vowel is perceptually tense, without concerning about which acoustic cues are responsible for the difference between tense vowels and lax ones. L1 English speakers' accentedness judgments on the L2 production [θik] could subsequently uncover whether vowel-tensing (e.g., using /i/ instad of /ɪ/) is perceptually foreign accented. Since the purpose of the current study was to investigate which phonetic patterns in L2 speech are perceptually foreign accented rather than how acoustic deviations affect perceived foreign-accentedness, we chose to utilize IPA transcriptions as a more efficient way of achieving our goal, rather than to investigate acoustic characteristics of the data

2.2 Prosodic Correlates of Foreign Accent

While research on segmental influences tends to disagree on which segmental cues are important, prosodic cues have also been found to be of vital importance in identifying foreign accent (e.g., Hahn, 2004; Kang, Rubin, and Pickering, 2010; Munro and Derwing, 2001; Zielinski, 2008). For example, prosodic cues such as intonation (Anderson-Hsieh, Johnson, and Koehler, 1992; Jilka, 2000), speaking rate (Munro and Derwing, 2001), lexical and phrasal stress (Kang, Rubin, and Pickering, 2010), and speech rhythm (White and Mattys, 2007) were often shown to affect foreign accent perception. The current study does not focus on the prosodic characteristics of L2 speech. However, it should be noted that some studies claimed that prosodic cues could outweigh segmental cues in accentedness perception (e.g., Magen, 1998; Munro and Derwing, 1995).

For example, Magen (1998) found that L2 English lexical and phrasal stress variations were perceived as more foreign-accented than L2 English segmental variations (e.g., vowel reduction, stop voicing, tense-lax alternation, using [ʃ] instead of [tʃ], and using [z] instead of [s]). Similar results were reported by Munro and Derwing (1995), who first asked linguists to determine the amount of L2 segmental and intonational variations contained in L2 English speech samples. L1 English listeners were then recruited to rate the foreign-accentedness of the L2 English speech samples. Their results show that L2 intonation correlated more with accentedness than did L2 segmental variations.

Since prosodic aspects of L2 speech potentially play a role in foreign-accentedness perception, it is, therefore, necessary for the current study to control for prosodic information while investigating

the role of segmental effects on foreign-accentedness perception. The specific method employed in the current study is discussed in Chapter 4.

2.3 Syllable Structure Correlates of Foreign Accent

As we have seen, most research on foreign accent perception has focused on the impact of vowels, consonants and prosody. Fewer studies have investigated the impact of changes in syllable structure on accentedness perception. L2 syllable production often involves some form of a simplification strategy, namely, segment epenthesis or segment deletion (Hansen, 2004; Sato, 1984). The substitution of a segment preserves the original syllable structure, while the addition or deletion of a segment changes the original syllable structure. The current study, therefore, considers segmental epenthesis and segment deletion as syllable structure variations rather than as segmental variations (Chapter 3).

Magen (1998) suggested that segment epenthesis is more salient than consonant variations in signaling foreign accentedness. Her study showed that epenthetic schwa was perceived as more accented than consonant feature changes (e.g., [tʃ] to [ʃ]). However, evidence is lacking as to whether segment deletion could also be indicative of foreign-accentedness. After all, a strategy such as obstruent coda deletion is also a prominent feature in L1 speech (Demuth, Culbertson, and Alter, 2006; Labov, 1997). Take t/d-deletion in English as an example. L1 English speakers are more likely to delete /t/ or /d/ when they are past tense morphemes (e.g., /d/ in "called," /t/ in "*packed*") than when they are part of the stem of a word (e.g., /d/ in "hold," /t/ in "*pact*") (Guy, 1991). By contrast, L2 speakers' t/d-deletion strategy does not seem to concern whether the /t/ or the /d/ is part of the stem of a word (Edwards, 2011; Hansen, 2004). Although there are indeed differences between deletion strategies in L1 and L2 speech productions, there is a paucity of evidence as to whether the differences affect foreign accent perception.

2.4 Accentedness Rankings of L2 Variations

While most studies have investigated only a few phonetic patterns in L2 speech, Magen (1998) and van den Doel (2006) compiled a list of L2 phonetic variants and directly compared their perceptual accentedness or "severity." In Magen (1998), two Spanish speakers each recorded 96 sentences in English, from which 56 phrases were selected for acoustic manipulation. For each phrase, Magen (1998) acoustically edited out one L2 variant (e.g., removed epenthetic schwa, lengthened VOT on [p^h], shortened vowel duration to create reduced vowels, removed a burst to change [tʃ] to [ʃ], resynthesized intonation contours to manipulate lexical and phrasal stress etc.), which would ideally make the altered phrases less accented than the original ones. Ten L1 English speakers provided their accentedness judgments on the synthesized phrases and their unaltered counterparts. By comparing judgment ratings of the altered and unaltered phrases, Magen (1998) found that lexical stress shifts, epenthetic schwas, vowel quality changes (e.g., [ʃɪp] becomes [ʃip][1]), and consonant manner of articulation changes (e.g., [tʃ] becomes [ʃ]) significantly affected accentedness perception, whereas plosive de-aspiration (i.e., VOT-shortening) did not.

While Magen (1998) mainly focused on Spanish speakers' L2 English productions, van den Doel (2006) focused on Dutch speakers' L2 English productions. To provide natural sounding stimuli, van den Doel (2006) asked L1 English speakers to mimic L2 phonetic variants that are common among Dutch speakers (e.g., [bɛd] becomes [bɛt]). He then placed these stimuli in carrier phrases (e.g., "*she lay in bed/bet for most of the day*") produced by the same L1 English speakers. He then asked L1 English listeners to first identify the L2 variants presented in each phrase, and then provide a severity rating on each L2 variant. The results showed that lexical stress shift and the uvularization of English /ɹ/ were judged to be the most severe among all variants (e.g., lexical stress shift, substitutions of /θ,ð/ by /t,d/, epenthetic [ə] in /lm/, de-aspiration of /t/, substitutions of /v/ by /w/, substitutions of /e/ for /æ/, yod-insertion in the word "*new*,"etc.). Although various consonant variants (e.g., VOT-shortening) and vowel variants (e.g., using /e/ instad of /æ/) were considered severe to L1 English listeners, consonant and vowel variants in general did not show any apparent

[1] According to Magen (1998), Spanish speakers' production of English lax vowel /ɪ/ contains a /i/ portion at the beginning of the vowel and a /ə/-like portion near the end. To make the L2 /ɪ/ more native-like, Magen removed the /i/ portion entirely, and then lengthened the /ə/-like portion

difference in severity.

Magen (1998) and van den Doel (2006) both studied a specific group of L2 English speakers, and both provided an accentedness or "severity" ranking for different types of phonetic patterns in L2 speech. They both found that lexical stress shift and vowel epenthesis are indicative of accentedness, but they seemed to disagree on whether plosive de-aspiration (i.e., VOT-shortening) affects accentedness perception. The two studies also applied different approaches to achieve experimental control. Magen (1998) resorted to acoustic manipulations, while van den Doel (2006) had L1 English speakers mimic L2 English variations. Both strategies have advantages and shortcomings. Acoustic manipulation can be quite precise in altering specific signals, but one might question the "naturalness" of the altered sound. L1 speakers' mimicry of L2 English variants might indeed achieve a certain degree of naturalness; however, it is debatable whether the mimicry is truly representative of L2 speech or not. Both Magen (1998) and van den Doel (2006) placed stimuli in carrier phrases. However, the phonological context of each target stimulus was not well controlled, raising question about whether phonological context affected accentedness judgments.

The current study aims to address potential problems in the previous literature by using unaltered L2 speech samples as stimuli and controlling for phonological contexts. Based on findings in the previous literature, the current study hypothesizes that various phonetic patterns in L2 speech do not carry equal weight in accentedness perception. Specifically, consonants are more important than vowels in accentedness perception. The rationale for this hypothesis could be discussed via theories of speech perception and lexical identification. The following section provides an overview of relevant theories and empirical findings.

2.5 Speech Perception Models

Speech perception is generally considered a process of mapping variable acoustic signals to linguistic representations (Holt and Lotto, 2010). The Native Language Magnet Model (NLM: Kuhl, 1991), for example, claims that infants gradually learn the perceptual prototypes for each L1 sound category by observing the distributional frequencies of speech sounds. The prototypes then function as a magnet to attract similar sounds, and together they form a sound category. This cluster of sounds

will not be easily discriminated from the prototype. Therefore, when one hears a sound that is near one of the L1 prototypes, the sound will be perceived as native-like.

The Perceptual Assimilation Model (PAM: Best, 1995) similarly suggests that there is a set of L1 sound categories to which L2 sounds may or may not be assimilated. PAM predicts that the accuracy of perceiving L2 sounds depends on how closely the sounds can be assimilated into existing L1 sound categories. Two predictions under the PAM framework might be relevant to the perception of foreign accent.

The first one is called Two-Category Assimilation (TC), which predicts that listeners can easily discriminate two sounds that belong to two separate L1 sound categories (e.g., /m/ vs. /n/); The second is called Category-Goodness Difference (CG) that predicts that when both sounds fit into the same L1 sound category but one is a better fit than the other (e.g., [t] vs. [t̪]), listeners are moderately good at discriminating the two sounds. In terms of foreign accent perception, TC can be used to predict phonemic alternations, while CG might explain sub-phonemic alternations.

Both NLM and PAM emphasize how the categorization of sounds affects speech sound discrimination. Plosive consonants, for example, are perceived categorically. Allophones of the same plosive consonant phoneme are, therefore, not easily discriminable. Consonant discrimination can thus be accurately predicted by consonant categorization (Liberman et al., 1957; Pisoni, 1973). However, discrimination does not always depend on categorization (Mirman, Holt, and McClelland, 2004; Repp, 1984). Vowel perception, for example, is often observed to be more continuous (Pisoni, 1973). Vowel discrimination (especially in steady-state vowels) is often easier than vowel categorization (Mirman, Holt, and McClelland, 2004). In other words, listeners might be able to detect the idiosyncratic differences between allophones of the same vowel phoneme but might not be able to tell whether the allophones belong to the same phoneme (Mirman, Holt, and McClelland, 2004). Such evidence can be found in Kronrod, Coppess, and Feldman (2012), who show that consonant discrimination relies more on category means in L1 speech (e.g., mean VOT for plosives, mean friction frequencies for fricatives). Vowel discrimination, on the other hand, does not solely rely on category means (e.g., mean formant frequencies). Studies using fMRI showed that the perceptions of consonants and vowels are indeed governed by two different neural mechanisms. The processing

of consonants is more left-lateralized than the processing of vowels (Altmann et al., 2014).

Due to the difference in perceptual strategies, vowel categorization is considered to be harder than consonant categorization. Altmann et al. (2014), for example, show that between-category consonants (e.g., /d/ vs. /b/) are more frequently categorized as separate phonemes than between-category vowels (e.g., /o/ vs. /a/). This finding, however, does not mean that between-category vowels are the same as within-category vowels. Between-category vowels are still more likely to be categorized into separate phonemes than within-category vowels (Altmann et al., 2014).

The aforementioned findings might explain why research on foreign accentedness often shows relatively consistent results for consonants, but conflicting results for vowels. The perception of consonants is categorical, and thus largely depends on L1 speaker norms. Listeners can detect the differences between between-category consonants (i.e., phonemes), while within-category consonants (i.e., allophones) are perceptually indistinguishable. The perception of vowels, on the other hand, is less categorical. L1 speakers might be able to detect differences between an L2 vowel and the target vowel regardless of whether the L2 vowel is categorically different from its L1 target.

Given the categorical perception of consonants, phonemic alternations of consonants (e.g., /b/ becomes /p/ in English) would lead to a higher degree of discrimination between the substituted (e.g., /p/) and the original sound (e.g., /b/) as a result of perceptual categorization. The phonemic alternations of vowels (e.g., /æ/ becomes /ɑ/) will similarly lead to a certain degree of discrimination. However, as discussed above, vowel discrimination does not necessarily entail vowel categorization. Differences between vowels are likely to be detectable even when the vowels are allophones of the same phoneme. Since L1 English listeners can detect the differences between vowel allophones, they could possibly consider differences between some L2 vowels and the L2 vowels' L1 counterparts as an allophonic difference. .

In addition to the relatively continuous perception of vowels, L1 English vowel phonemes exhibit considerable within-speaker and between-speaker variations (Hillenbrand et al., 1995; Peterson and Barney, 1952). Peterson and Barney (1952) investigated 78 L1 American English speakers' productions of ten English vowel phonemes. Results show that the ten vowel phonemes overlap considerably in the F1-F2 space.

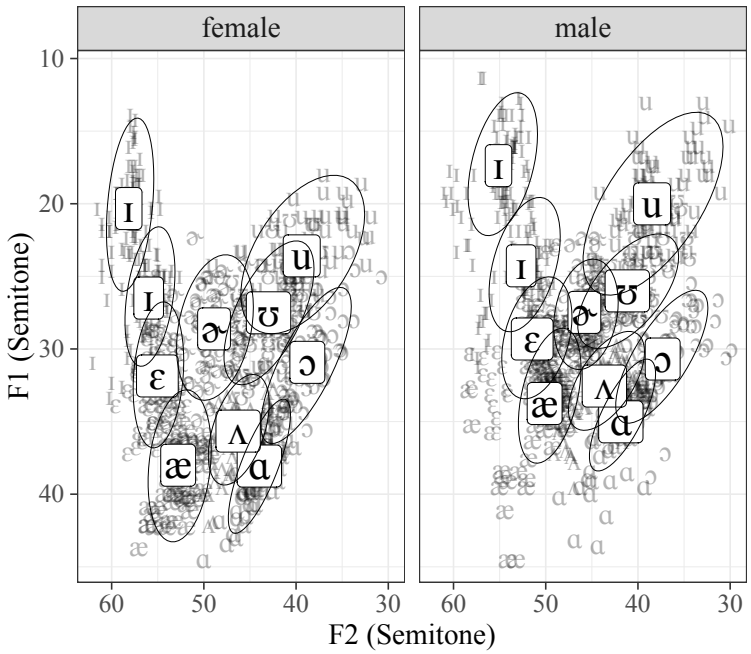

Figure 2.1: Vowel Space of 10 American English Vowel Phonemes

The data demonstrated in 2.1 are from Peterson and Barney (1952). F1 and F2 values of the vowels were converted to semitones relative to 100 Hertz (Hz). The gray phonetic symbols represent individual vowel productions by 78 L1 American English speakers. The ten black phonetic symbols represent mean F1 and F2 values. Male and female participants were drawn separately. Following Peterson and Barney (1952)'s practice, a closed loop for each vowel was drawn to enclose 90% of the observations.

As shown in Figure 2.1, F1-F2 spaces of some vowels encroach on the spaces of other vowels. This phenomenon makes it possible for some vowel productions to be considered allophonic variants of more than one vowel phoneme. It is, thus, possible that the encroachment of vowel space on the F1-F2 plane could potentially obfuscate the perceptual categorization of vowels. Such a claim was supported by Ladefoged (1989) who showed that the same test pronunciation /bɪt/ could be perceived as "bit" by some L1 English listeners, but as "bet" by other L1 English listeners.

The current study investigates the effect of categorizability on accentedness judgment. If accentedness perception depends on how easily an L2 speech sound can be categorized into its target L1

sound category, then consonant changes (e.g., using [ʃ] instead of [tʃ]) might be more accented than vowel changes (e.g., using [ʌ] instead of [æ]), because consonants are more categorizable. Given the relatively higher categorizability of consonants, it is expected that listeners will more likely agree on which consonant phoneme they hear in an utterance. Disagreements might emerge when vowels are concerned. If the categorizability of speech sounds participates in accentedness judgment, then judgments on consonants produced by L2 speakers could be more consistent than judgments on vowels produced by L2 speakers.

2.6 Lexical Identification

Another possible difference between consonant and vowels lies in their respective role in lexical identification. Previous research often claims that consonants are more important than vowels in lexical identification (Nespor, Peña, and Mehler, 2003). Studies in lexical identification often attributes the difference between consonants and vowels to their phonological-distributional properties (Nespor, Peña, and Mehler, 2003). They claim that vowels are more variable than consonants in continuous speech. Listeners who are accustomed to contextual variabilities of vowels would therefore choose to rely on consonants to identify words (Cutler et al., 2000). If lexical identification participates in one's judgment of accentedness, then perhaps consonant changes would be considered more accented, because lexical identification depends more on consonants than vowels.

Since vowels are generally longer and louder than consonants (Ladefoged and Maddieson, 1996), one would intuitively assume that vowels are more salient to listeners and should consequently be more reliable in lexical identification. However, results from previous empirical studies have shown just the opposite. For example, participants of (Bonatti et al., 2005) preferred identifying words in continuous speech using the transitional probability (TP) between consonants, rather than TP between vowels. TPs of speech sounds are defined as the conditional probabilities of sound B, given A. Intuitively, the TP between A and B could be interpreted as the probability for B to follow A in the same word (e.g., how probable it is for /s/ to follow /k/ in the same word).

In Bonatti et al. (2005), participants first listened to nonce words of an artificial language (e.g., /mulitɛ̃/, /mylɔ̃ta/, /budikɛ̃/, /bydɔ̃ka/). The words were designed to keep TPs between consonants

and TPs between vowels of a word constant (e.g., consonants /m, l, t/ and /b,d, k/ always occurring in the same order in a word, vowels /u, i, ɛ̃/ and /y, ɔ̃, a/a row in a word).

After being exposed to the nonce words, participants heard two more types of nonce words. The first type, called the CT words, changed vowel sequences, but kept consonant sequences the same as the words heard during the exposure phase (e.g., /mɔ̃latɑ̃/, /bidɛ̃ky/). The second type, called the VT words, kept vowel sequences the same as before but changed consonant sequences (e.g., /dukimɛ̃/, /ʁygɔ̃ma/). Participants judged the CT words to be more similar to words they heard during the exposure phase, suggesting that consonant information was more salient than vowel information to the participants.

In van Ooijen (1996)'s word reconstruction test, participants were asked to change a nonce word (e.g., *kebra*, *eltimate*) to a real word by changing either the first consonant (e.g., turn *kebra* to *zebra*, turn *eltimate* to *estimate*) or the first vowel (e.g., turn *kebra* to *cobra*, turn *eltimate* to *ultimate*). Participants overwhelmingly preferred changing the vowels, rather than changing the consonants. When specifically asked to change only the consonant or only the vowel, participants were significantly better at performing vowel changes than consonant changes. These results were replicated by (Cutler et al., 2000), who performed experiments on Dutch and Spanish speakers, using Dutch or Spanish words as stimuli. Cutler et al. (2000) further showed that the preference for vowel alternations was unrelated to the consonant-to-vowel ratio in a specific language. Results from these studies show that listeners, regardless of their linguistic background, tend to tolerate vowel changes more so than consonant changes.

Bonatti et al. (2005) and van Ooijen (1996)'s results have significant implications in foreign accent research. Thus, if consonant information is truly more important than vowels in lexical identification, then consonant change in L2 speech would be considered a more severe departure from its L1 target, which could potentially lead to a higher degree of foreign accentedness. However, if accentedness judgment does not concern lexical information, then consonant and vowel changes might not differ in their accentedness. To evaluate whether lexical identification participates in accentedness judgment, lexical information of an L2 utterance needs to be taken into consideration, in addition to phonetic and phonological information. Chapter 6 discusses this issue in detail.

2.7 Summary

Previous studies have suggested that both the temporal (e.g., VOTs of plosives) and spectral (e.g., formants of liquids) aspects of consonants have an effect on accentedness ratings. The findings on vow- els were mixed, with some studies showing the effects of both temporal (e.g., vowel duration) and spectral aspects (e.g., F1 and F2) of vowels (McCullough, 2013), while others showing only the spectral effects (Chan, Hall, and Assgari, 2016). Among the studies that show spectral effects, some claim that the degree of spectral deviation from native speaker norms can predict accentedness (McCullough, 2013; Wayland, 1997), whereas others argue that the spectral overlapping of vowel categories is the key in foreign accent perception (Chan, Hall, and Assgari, 2016; Sidaras, Alexander, and Nygaard, 2009). Despite the disagreement, previous literature show that accentedness is affected by some phonetic feature of vowels, be it spectral overlapping of vowel categories or spectral deviation from a native speaker's norm. Among the research on foreign accent perception, the perception of L2 syllables has not been well studied. Although there is some evidence of the relative importance of segment epenthesis, questions remain as to whether segment deletion correlates with accentedness.

The current study aims to address the potential problems in previous research by obtaining stimuli that are both natural and representative of L2 speakers from various language backgrounds. Instead of acoustic manipulation, the current study opted to pick stimuli that were already identified as containing phonetic features that are common to L2 English speakers. The current study also controls for prosody in the least intrusive manner by employing a Dynamic Time Warping method. Details of stimuli selection and this Dynamic Time Warping method are discussed in Chapter 4.

Many previous studies have focused on only a few types of phonetic patterns in L2 speech The current study investigates the perceptual accentedness with a larger variety of phonetic patterns. The results provide a more detailed understanding of how different types of phonetic patterns are weighted in accentedness perception. The results are discussed in Chapters 4 and 5.

The current study hypothesizes that consonants and vowels do not contribute to accentedness to the same degree. Since consonants and vowels differ in their perceptual categorizability, it is possible that phonemic alternations of consonants and vowels are perceived differently, which could in turn

affect their perceptual accentedness. Since consonants and vowels were often observed by previous literature to function differently in lexical identification, it is possible that consonant and vowel changes exhibit different degrees of foreign accentedness. Based on previous research on L2 syllable structures, the current study hypothesizes that segment epenthesis is perceptually more accented than segment deletion because segment deletion is sometimes allowed in L1 speech, while segment epenthesis rarely occurs. The current study empirically examines these hypotheses in Chapters 4 and 5.

Chapter 3: Stimuli Selection

3.1 Introduction

The current study aims to investigate native (L1) English speakers' accentedness judgment on non-native (L2) English speech samples. To this end, 100 L2 audio speech samples served as stimuli for two perception studies. These two studies are discussed in Chapters 4 and 5. This chapter describes how the stimuli were selected.

The 100 L2 speech samples were selected from the Speech Accent Archive (SAA: Weinberger, 2019). Audio samples in the SAA are phonetically transcribed by trained phoneticians using the International Phonetic Alphabet (i.e., IPA). These transcriptions reflect the transcribers' perception. The 100 L2 speech samples were selected primarily based on their respective IPA transcriptions. Although care has been taken by the SAA personnel in providing relatively reliable transcriptions, inaccuracies could still exist. Therefore, the current study conducted acoustic analysis on the samples to further evaluate the reliability of the IPA transcriptions.

In addition to the 100 L2 speech samples, speech samples from 50 L1 speakers of American English were also extracted from the SAA to evaluate the acoustic difference between L1 and L2 speech samples. Mean acoustic measurements of these 50 L1 speech samples were used as an approximation for native speaker norms. If the acoustic difference between an L2 speech segment and its native speaker norm could be captured by the IPA transcription for the L2 speech segment, then the IPA transcription is considered reliable and is used in the current study.

Some of the acoustic measurements described in this chapter were further utilized in Experiment 2 (Chapter 5) to see how they correlate with perceived accentedness. It should be noted that acoustic correlates of a phoneme are numerous and could vary depending on adjacent phonological contexts and sometimes extra-linguistic factors such as age (Holt and Lotto, 2010). Consequently, a few acoustic measurements cannot fully capture all of the features of a phoneme. To avoid the complexity

of acoustics-to-phoneme mapping, the current study selected the stimuli primarily based on the IPA transcriptions, which are reflective of perception. The acoustic analysis described in this chapter serves mainly to verify the IPA transcriptions. The following section describes how stimuli selection and acoustic verification were carried out.

3.2 Corpus

The 100 L2 stimuli are short two-word speech samples selected from the SAA. The SAA, as of June 2019, consists of 2786 audio speech samples produced by both L1 and L2 English speakers. All samples consist of speakers reading the sample paragraph given below:

The "Stella" Passage:

Please call Stella. Ask her to bring these things with her from the store. Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake, and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station.

The speakers also provided their demographic information (e.g., age, gender, native language, age of English onset, learning style, current and former residence, etc.), which is publicly available on the SAA website. The speech samples were all digitally recorded by trained experimenters in the acoustic lab at George Mason University or somewhere else where the speakers and the experimenters deemed suitable. Only CD-quality (16-bit/44kHz) recordings are included in the archive. Low quality recordings are not accepted (Weinberger and Kunath, 2011).

The audio samples were narrowly transcribed using the IPA by at least three transcribers. The transcribers were phonetically trained graduate students at George Mason University. The IPA transcriptions were reached by the consensus of at least three people. All transcriptions were vetted and made available online by the administrator (i.e., Dr. Steven Weinberger) of the SAA.

It should be mentioned that speech samples in the SAA were not transcribed by the same group of transcribers. Although most of the transcribers are L1 speakers of American English, some of them are international students whose native language is not English. Dr. Weinberger administered

quality control by training all the transcribers through a semester-long, graduate-level phonetics class. Transcribers were instructed to follow the same guideline when transcribing the speech samples. To improve accuracy and resolve disagreements among the transcribers, speech processing software such as PRAAT (Boersma and Weenink, 2015) or Audacity (Mazzoni and Dannenberg, 2000) were often used to analyze benchmark acoustic signals. Around 1300 speech samples have been transcribed and are available online. An example is given below:

[pʰliːz kɔːl stɛla ɛsk ɝtu bɹĩŋ ðɪ θĩŋs wɪθ hɜɹ fɹõm θə stɔːɹ sɪks ĭspũːnz ɔf fɹɛʃ snoʊ pʰiːz faɪf θɪk slɛːbz̩ ɔv bluː tʃĩːz ẽn meɪbi ə snẽːk fɔɹ hɛɹ bɹʌðəɹ bɑːb̃ wi ɔlso niq̇ ə smɔːl plȩːstɪk sneɪk ẽn ə bik t̪ɔɪ fɹɑːg fɔr ðə kɪːts ʃi kẽn skuːp̪ ðiːz θĩnks ĩntu θɹɪ ɹɪd bɛgz ẽn wi wɪl goʊ miːt hɛɹ wẽnɪzdeɪ æ d̪ə tɹeĩn steɪʃə̃n] (SAA: arabic23)

The vetted IPA transcriptions in the SAA were previously utilized by researchers in computational linguistics (e.g., Frost, 2013; Minematsu et al., 2014) and second language acquisition (e.g., Klein, 2011). Recently, Weinberger et al. (2019) recruited an additional 67 phonetically trained transcribers to transcribe a selection of audio clips from the archive. The results show that 72% of the 67 participants' transcriptions matched the vetted ones, lending support to the reliability of the vetted transcriptions.

3.3 Types of Stimuli

The 100 L2 speech samples for the current study were selected based on the vetted transcriptions available in the SAA. These samples were further grouped into four types based how they differ from their L1 target productions. This section describes the determination of L1 target productions, and the four types of stimuli used by the current study.

3.3.1 L1 Target Productions

The L1 target productions were defined as the most common L1 productions. Such a treatment was based on the assumption that L1 American English listeners would all be familiar with the most

common L1 productions and consequently rate them as native-like. To find the most common L1 productions, the current study surveyed productions from 100 L1 speakers of American English in the SAA (See Appendix A for demographic information). The most common productions among these speakers were considered L1 target productions. For example, 90% of the L1 productions for the word "*thick*" are [θɪk]. [θɪk] was therefore chosen as the L1 target production for "*thick*." IPA transcriptions of these L1 American English speakers were also used to build a computational model to approximate English phonetic and phonological grammar. Chapter 6 discusses the model in detail.

For the two perception experiments, the current study selected 100 L2 speech samples as stimuli. L2 speech samples that were transcribed the same as their L1 target production (e.g., [θɪk] for "*thick*") were termed the "match" stimuli, meaning that the L2 productions match their L1 target productions. L2 productions that do not match their L1 target productions were termed the "mismatch" stimuli. Some of the mismatch stimuli are not unique to L2 speakers of English. For example, some L1 speakers pronounce "*thick*" as [tɪk] according to data on the SAA. Since [tɪk] does not match the L1 target [θɪk], [tɪk] was nevertheless termed as a mismatch stimulus. Presumably, the mismatch stimuli that are observed in L1 speech data would be judged as less accented. The "mismatch" that are not observed in L1 speech data (e.g., pronouncing "*ask*" as [æskə]) would be judged as more accented.

3.3.2 Four Types of Stimuli

The mismatch stimuli were further divided into three groups based on three types of mismatches, namely, stimuli with consonant mismatches (e.g., [tɪk] for "*thick*"), stimuli with vowel mismatches (e.g., [θik] for "*thick*"), and stimuli with syllable structure mismatches (e.g., [faɪvə] for "*five*"). Therefore, the 100 stimuli were divided into four types (i.e., the match type and the three types of mismatches).

Table 3.1 illustrates the four types of stimuli. IPA transcriptions in the match stimuli column represent the most common L1 pronunciations for the five phonological contexts. 25 of the 100 L2 stimuli were transcribed the same as the IPA transcriptions listed in the match column. These 25 L2

speech samples were called the match stimuli [1]. The remaining 75 are mismatch stimuli. They differ from the match stimuli by only one phonetic element. For example, stimulus [faɪv tɪk] differs from its corresponding match stimulus [faɪv θɪk] by only one consonant (i.e., [θ]→[t]). Stimulus [faɪv tɪk] therefore contains only one consonant mismatch. Among the 75 mismatch stimuli, 25 contain only one consonant mismatch (five stimuli for each context), 25 contain only one vowel mismatch (five stimuli for each context), and 25 contain only one syllable structure mismatch (five stimuli for each context).

Table 3.1: Types of Stimuli

Context	Match	Consonant Mismatch	Vowel Mismatch	Syllable Mismatch
please call	[pʰliz kʰɑl]	[**p**liz kʰɑl]	[pʰliz kʰ**o**l]	[pʰ**ə**liz kʰɑlᵛ]
ask her	[æsk (h)əɹ]	[æsk hə**r**]	[**ɑ**sk həɹ]	[æs_ həɹ]
six spoons	[sɪks spunz]	[sɪks spun**ʃ**]	[s**i**ks spunz]	[sɪks **ə**spunz]
five thick	[faɪv θɪk]	[faɪv **t**ɪk]	[f**a**v θɪk]	[faɪv**ə** θɪk]
small plastic	[smɑl pʰlæstɪk]	[smɑ**l̩** pʰlæstɪk]	[smɑl pʰlæst**i**k]	[smɑl pʰlæs_ɪk]

3.4 Acoustic Comparisons

The 100 L2 stimuli represented 11 types of consonant mismatches, five types of vowel mismatches and two types of syllable structure mismatches. The stimuli were chosen mainly based on the IPA transcriptions available in the SAA. In order to verify the IPA transcriptions, acoustic analyses were carried out to compare benchmark measurements between the L1 and L2 speech samples. The aim of the following acoustic analysis is to verify whether acoustic differences between an L2 stimulus and its corresponding L1 speech samples can be captured by the IPA transcription of the L2 stimulus. For example, if an L2 segment was transcribed as an aspirated [pʰ], while its corresponding L1 productions are the non-aspirated [p]s. The [ʰ] symbols suggests that the L2 segment has a longer voice onset time (VOT) than the L1 [p]s. To verify whether the [pʰ] was correctly transcribed, the

[1] [æsk (h)əɹ] in the match column means that both [æsk əɹ] and [æsk həɹ] are equally common in the SAA.

current study measures the duration of the L2 VOT and the mean VOT of L1 [p]s. If the L2 VOT is indeed longer than the mean L1 VOT, then the current study accepts [p^h] as a correct transcription for the L2 segment.

L1 English speech samples from the SAA were selected for the determination of what might constitute native speaker pronunciation norms. The speech samples were produced by 50 L1 speakers of American English. They were from 21 states in the continental U.S., ages from 18 to 79 (M=39.67, SD=16.62). 25 of them are male, the other 25 are female. Detailed demographic information of these 50 L1 speakers is listed in Appendix A.

A large body of previous research has demonstrated that phonemic categories are distinguished by multiple acoustic cues (e.g., Lisker, 1986; Toscano and McMurray, 2010). For example, VOT, pitch of the following vowel, and other 14 kinds of acoustic cues could be responsible for the distinction between /pa/ and /ba/ (Lisker, 1986). The current study opts to focus on the primary and most robust cues as identified by previous speech perception research, while fully acknowledging that other phonetic cues could also affect speech perception.

3.4.1 Segmentation

All the L1 and L2 speech samples were first segmented. The intensity of the speech samples was normalized to 75dB using PRAAT. Initial phoneme segmentation of the speech samples was performed with the Montreal Forced Aligner (McAuliffe et al., 2017). This is a newly developed neural network-based aligner that is comparable to human annotators and provides better performance than the traditional Penn Phonetics Lab Forced Aligner (Yuan and Liberman, 2008). Errors or inaccuracies discovered in the auto-segmentation were manually corrected. Relevant acoustic measurements were extracted via a PRAAT script for further analysis. Figure 3.1 illustrates the results of the automated alignment, where the top row displays information of the spectrogram and formants and the second row from the top displays pitch (dotted line) and intensity (solid line) contours. The third row from the top represents phone boundaries using the Arpabet symbols and the bottom row shows word boundaries.

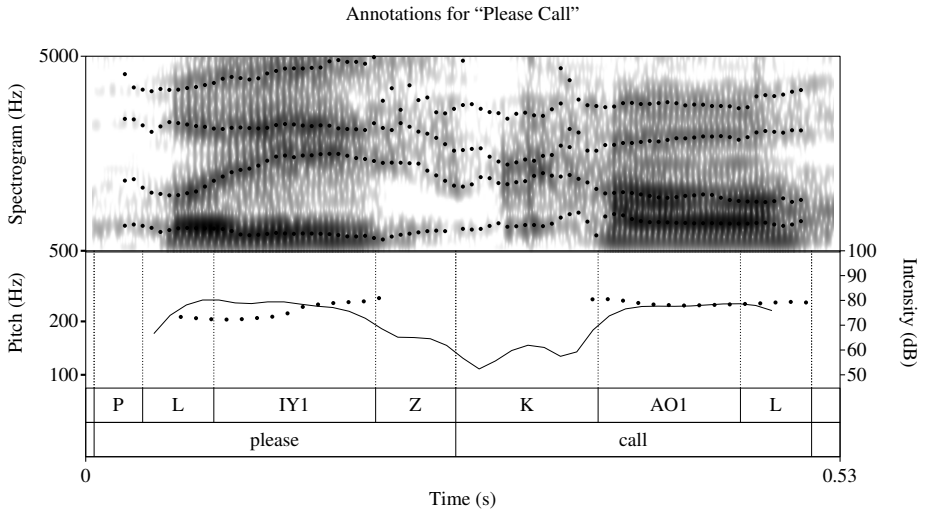

Figure 3.1: Illustration of Auto-segmentation

3.4.2 Plosives

Six L2 speech samples selected by the current study contained plosive-related mismatches as indicated by their respective IPA transcriptions. The following describes how the mismatches were verified by acoustic measurements. Duration of voice onset time (VOT) was used as the benchmark acoustic measurement for plosives. VOT duration of L2 segments and mean VOT duration of L1 segments were measured.

VOT is defined as the interval between the onset of a plosive burst and the onset of the following vocalic onset. There are numerous claims that VOT directly affects accentedness perception (Major, 1987; Riney and Takagi, 1999). Six L2 stimuli were therefore selected to verify these claims made in previous research. Following the practice of Chodroff and Wilson (2017), the beginning of the VOT was placed at the beginning of a plosive burst release; and the endpoint was placed at the beginning of periodicity in the waveform or a visible pitch track, whichever came first. VOT labeling was initially achieved in PRAAT with the AutoVOT plugin (Keshet, Sonderegger, and Knowles, 2014). Labeling errors were manually corrected in PRAAT based on waveforms and spectrograms.

Contexts "*please call*," "*small plastic*" and "*six spoons*" were used for the investigation of VOT-related consonant mismatches. Six L2 speech samples that were identified as having VOT-related mismatches were extracted for analysis. Two speech samples involved VOT-shortening on [k^h] in the word "call," one speech sample involves VOT-shortening on [p^h] in the word "please," one speech sample involves VOT-shortening on [p^h] in the word "*plastic*," two speech sample involve VOT-lengthening on [p] in the word "*spoons*." In addition to the L2 speech samples, VOTs of 50 L1 speech samples were measured for comparison.

Table 3.2 illustrates the type of VOT-related consonant mismatches and the contexts where they occurred. Every row in Table 3.2 represents the VOT of a plosive segment. The mismatch column lists the type of VOT-related mismatches and the segments involved. The L2 VOT column contains durational measurements of the six L2 VOTs. L2 VOTs were converted to z-scores with regard to L1 English VOT means and standard deviations (SD). A z-score represents how many standard deviations an L2 VOT mean is from the mean L1 VOTs. A positive z-score means an L2 VOT is longer than the mean L1 VOT, while a negative z-score means an L2 VOT is shorter than the mean L1 VOT.

For example, the first row of Table 3.2 shows that an L2 stimulus "*please call*" involves the de-aspiration of [k^h] in the word "call." The L2 segment [k] has a VOT of 33.01 milliseconds (ms). Calculation based on 50 L1 American English productions of "*please call*" shows that the mean VOT of L1 [k^h]s is 52.78 ms with a standard deviation of 13.71 ms. The z-score shows that the L2 VOT is 1.44 standard deviations below the L1 mean.

Table 3.2: L1 and L2 VOT Comparisons

Contexts	Mismatches	L2 VOT (ms)	L1 (English) VOT (ms)	Z
please call	[k^h]→[k]	33.01	M=52.78; SD=13.71	-1.44
please call	[k^h]→[k]	21.26	M=52.78; SD=13.71	-2.30
please call	[p^h]→[p]	10.05	M=62.50; SD=18.06	-2.91
small plastic	[p^h]→[p]	20.86	M=62.84; SD=15.53	-2.70
six spoons	[p]→[p^h]	45.33	M=14.46; SD=7.15	+4.32
six spoons	[p]→[p^h]	68.53	M=14.46; SD=7.15	+7.56

As shown in Table 3.2, the shortened VOTs are indeed shorter than the L1 means, while the lengthened VOTs are indeed longer than the L1 means. Therefore, VOT differences between the L1 and L2 speech samples were successfully captured by the IPA transcriptions of the L2 speech samples. These L2 speech samples were therefore chosen to represent VOT-related phonetic mismatches in L2 speech.

3.4.3 Fricatives

Seven L2 stimuli were selected to investigate the perceptual accentedness of fricative-related mismatches. Four stimuli involve replacing L1 fricatives with other fricatives (e.g., [z]→[s], [θ]→[f]). Three stimuli involve replacing the interdental fricative /θ/ with /t/. The seven stimuli are listed in Table 3.3. Analysis in this session concerns three types of benchmark acoustic signals, namely, Center of Gravity, pitch context, and noise ratio. These acoustic signals were selected to approximate differences in place and manner of articulation.

Center of Gravity

Previous research on the acoustic correlates of English fricatives has discovered that the Center of Gravity (COG) of fricatives is a reliable cue for place of articulation (Jongman, Wayland, and Wong, 2000). COG is a measurement of energy concentration. Energy of a speech sound, as measured by amplitude, could be concentrated in either the higher frequencies or the lower frequencies of the sound. A smaller COG implies that the energy of a sound is concentrated in the lower frequencies.

As a measurement for place of articulation, COG value decreases as place of articulation moves further back in the oral cavity. For example, alveolar fricatives (e.g., /s/ and /z/) have lower COGs than dental fricatives (e.g., /θ/ and /ð/), whose COGs are lower than labiodental fricatives (e.g., /f/ and /v/) (Jongman et al., 2000). Although COG is traditionally thought of as a cue for place of articulation, Jongman, Wayland, and Wong (2000) reported that COG is also a good indicator for voicing, with voiceless English fricatives having significantly higher COGs than their voiced counterparts. Table 3.3 lists the COGs of segments in the seven L2 and L1 stimuli selected by the current study.

Table 3.3: L1 and L2 COG Comparisons

Contexts	Mismatches	L2 COG (Semitone)	L1 COG (Semitone)	Z
please call	[z]→[s]	77.68	M=71.14; SD=5.31	+1.23
small plastic	[s]→[z]	64.39	M=71.30; SD=7.28	-0.95
five thick	[θ]→[f]	73.77	M=64.37; SD=11.40	+0.82
six spoons	[z]→[ʃ]	71.10	M=64.56; SD=8.86	+0.73
five thick	[θ]→[t̪]	59.63	M=64.37; SD=11.40	-0.42
five thick	[θ]→[t̪]	63.27	M=64.37; SD=11.40	-0.10
five thick	[θ]→[t̪]	42.63	M=64.86; SD=9.71	-2.29

Every row of Table 3.3 shows the COGs of an L2 segment and the corresponding mean L1 COG value. For example, the first L2 stimulus is "*please call*." The coda of "*please*" was transcribed as [s]. The L2 production therefore involves the devoicing of the [z] in "*please*" (i.e., [z]→[s]). The COG of the L2 segment [s] is 77.68 semitones, while its corresponding L1 segment [z] has a mean COG of 71.14 semitones with a standard deviation of 5.31. The L2 segment [s] is thus 1.23 standard deviations above the mean. This result is consistent with previous claims that COGs of voiceless fricatives are higher than COGs of their voiced counterparts. Therefore, the [s] was accepted by the current study as a correct transcription.

The second L2 stimulus "*small plastic*" replaced the [s] in "*small*" with a voiced [z]. The COG of this L2 segment [z] is lower than the mean COG of the corresponding L1 [s]s. This finding is also consistent with previous research, which showed that voiced fricatives have lower COGs than voiceless fricatives (Jongman, Wayland, and Wong, 2000). Acoustic measurements for the third stimulus show that the L2 segment [f] has a higher COG than its L1 target segment /θ/, which is consistent with previous findings that COG value increases as the place of articulation moves further front in the oral cavity (Jongman, Wayland, and Wong, 2000). Given these results, the IPA transcriptions for the first three stimuli in Table 3.3 were considered accurate by the current study. The difference between the remaining four L2 stimuli and their corresponding L1 productions cannot be explained via the COG measurement. The following section discusses two additional acoustic measurements that could capture the difference between the remaining four L2 stimuli and their

corresponding L1 productions.

Pitch Context

The fourth L2 stimulus "*six spoons*" changed the final consonant in "*spoons*" from [z] to [ʃ]. The COG of the voiceless L2 segment [ʃ] is higher than the mean COG of its target L1 segment [z]. Previous research claimed that the COG of the post-alveolar fricative [ʃ] should be lower than both alveolar fricatives /s/ and /z/ (Jongman, Wayland, and Wong, 2000). Based on the COG measurement, [ʃ] is perhaps not a correct transcription. However, acoustic correlates of English phonemes are multidimensional. COG values alone cannot explain why the final consonant of "*spoons*" was transcribed as [ʃ].

In addition to COG values. the perception of /s/ and /ʃ/ could be affected by the pitch values of the preceding segments (Niebuhr, 2017). Specifically, /s/ is more likely to be perceived as /ʃ/ when its preceding segment carries a higher pitch. In other words, the final consonant in "*spoons*" could be an /s/, but was perceived and transcribed as [ʃ] because the penultimate segment /n/ carries a relatively higher pitch. To investigate whether the preceding pitch context of the final consonant in "*spoons*" potentially affected transcribers' perceptions, pitch values of the /n/s were extracted in PRAAT. Pitch values were extracted at the location of the energy peak in the amplitude spectrum, using methods described in De Jong and Wempe (2009). Pitch values were converted to semitones relative to 100 Hertz to approximate the non-linear mapping between Hertz values and human perception.

The L2 stimulus "*six spoons*" was produced by a male Lamaholot speaker. The [n] of the L2 stimulus carries a pitch of 20.98 semitones. To calculate pitch values of the natively produced /n/s, productions of 25 male L1 American English speakers were similarly processed in PRAAT. Results showed L1 male speakers' productions of the /n/ carry a mean pitch of 14.93 semitones, with the standard deviation of 3.27 semitones. The L2 /n/ thus indeed carries a relatively higher pitch than mean pitch values of the L1 /n/s. Therefore, there is a reason to believe that the pitch value of the penultimate consonant [n] in the L2 production of "*spoons*" affected transcribers' perception of the final consonant. This finding does not necessarily imply that the IPA transcriptions of the L2 stimulus are incorrect, but affirms the fact that acoustic correlates of phonemes are multidimensional.

Therefore, the current study accepted the [ʃ] as a correct transcription.

Noise Ratio

In addition to the four stimuli analyzed above, there are three stimuli that involve the replacement of the dental fricative /θ/ with a dentalized [t̪]. As discussed above, COGs increase as the place of articulation moves farther front in the oral cavity. Therefore, dental [θ]s should have higher COGs than [t̪]s. Table 3.3 shows that the mean L1 COG of [θ]s is indeed higher than COGs of the [t̪]s. However, the difference between /θ/ and /t/ lies primarily on their respective manners of articulation, which is not captured by COG measurements. To inspect whether the [t̪]s are reliable transcriptions, the current study opted to use noise ratio to investigate the manner difference between fricative and plosive consonants.

Plosive consonants, such as the alveolar /t/, consist of a silent closure interval, which is followed by a frication noise burst and an interval of aspiration noise. Fricative consonants, such as /θ/, also consist of a silent closure interval, which is followed by a period of friction noise. The longer the closure interval, the more likely a fricative is perceived as an affricate (Dorman, Raphael, and Isenberg, 1980). The shorter the duration of friction noise, the more likely a fricative is perceived as a plosive (Dorman, Raphael, and Isenberg, 1980).

To investigate durational measurements of the American English /θ/, the same 50 American English speakers' productions of "*five thick*" were used. Male and female samples were analyzed separately, because frication noise of females tends to be shorter than males (Jongman, Wayland, and Wong, 2000).

A possible confounding factor of durational measurements is speech rate. Intervals of closure and frication noise might be shortened in fast speech. In slow speech, the intervals could be lengthened. To control for speech rate, noise ratios were calculated. Noise ratios were defined as the ratio of fricative noise duration over the duration of the whole word (Jongman, Wayland, and Wong, 2000). Word duration was defined as the interval between the onset of the frication to the end of the word. Closure intervals were not included in the total duration of a word, because the L1 and L2 segments are word initial. It is difficult to distinguish closure intervals of word-initial segments

from speech pauses between words. Tabel 3.4 illustrates the noise ratios of the L1 and L2 segments.

Table 3.4: L1 and L2 Noise Ratio Comparisons

Phrases	Gender	Mismatches	L2 Noise Ratio	L1 Noise Ratio
five thick	male	[θ]→[t̪]	0.17	0.25
five thick	male	[θ]→[t̪]	0.15	0.25
five thick	female	[θ]→[t̪]	0.16	0.20

The mean noise ratio for male L1 speakers is 0.25 (SD=0.06), and the mean noise ratio for female L1 speakers is 0.20 (SD=0.08). For the three L2 stimuli that replaced /θ/ with [t] or [t̪], noise durations were defined as the duration of the release burst and the aspiration noise. Word durations were defined as the interval from the onset of the release burst to the end of a word. As shown in Table 3.4, the noise ratios for [t̪]s are 0.17, 0.15, and 0.16 respectively. In other words, L1 [θ]s, in general, have a longer frication noise duration than the noise durations (i.e. burst and aspiration noise) of the three corresponding L2 segments. Since shorter noise duration of a segment increases the chance for it to be perceived as a plosive consonant (Jongman, 1989), there is reason to believe that the three L2 segments were transcribed correctly. These L2 stimuli were therefore accepted by the current study as representatives of /θ/-stopping (i.e., /θ/→/t/).

3.4.4 Liquids

In addition to plosives and fricatives, the current study also aims to investigate the accentedness of liquid productions in L2 speech. L2 speech involving the alternation between /ɹ/ and /l/ was often perceived by L1 American English listeners as very accented (Riney, Takagi, and Inutsuka, 2005). Eight L2 speech samples were selected to investigate the accentedness of syllable final /l/s and /ɹ/s. Based on findings in previous research, the degree of /l/-velarization was approximated by the difference between F1 and F2 (Riney, Takagi, and Inutsuka, 2005) values. The degree of /ɹ/-rhoticity was approximated by the difference between F2 and the third formant frequencies (F3) (Ohala and Ohala, 2001). The F1, F2, and F2 values were extracted at the energy peak in the amplitude spectrum

of the liquids.The following section discusses the analysis on L1 and L2 liquid productions in the phonological contexts of "*small plastic*" and "*ask her*."

Formant Information of L1 Liquids

Previous acoustic research has shown that the velarization of the English word-final /l/ causes an increase in F1 and a decrease in F2, with respect to non-velarized /l/s (Riney, Takagi, and Inutsuka, 2005). Previous literature on world-final /ɹ/ often associates rhoticity with the lowering of F3. In addition to the English /ɹ/s, low F3 for rhoticity was shown to hold for Malayalam rhotic trills, Toda trills, Tamil retroflex /ɭ/ and Hindi retroflex plosives (Ohala and Ohala, 2001). Indeed, the difference between F2 and F3 has been recommended as a reliable acoustic cue for automated measurement of rhotics (Campbell et al., 2018). The current study therefore operationalized rhoticity by taking the difference between F2 and F3.

The labeling of liquid segments was achieved via the Montreal Forced Aligner with manual adjustments afterwards. The beginning of a word-final liquid was set at the start of the F2 transition, while the end of a word-final liquid was set at the beginning of a pause between words or the following segment (whichever occurred first). F1, F2 and F3 values of each liquid were then extracted with PRAAT at the location of the energy peak in the amplitude spectrum, using methods described in (De Jong and Wempe, 2009). Formant frequencies were then converted to semitones relative to 100 Hertz.

Figure 3.2 demonstrates the spectral information of word-final /l/s and /ɹ/s produced by the 50 native speakers of American English, where the bold phonetic symbols represent the means and the shaded areas represent one standard deviation around the means. The small gray symbols represent productions of the 50 native speakers of American English. Male and female productions were presented separately, because male voice frequencies are generally lower than female voice frequencies. In general, the /ɹ/s show a relatively smaller F3-F2 difference, demonstrating that the English word-final /ɹ/s are more rhotic than word-final /l/s. The word-final /l/s show a relatively small F2-F1 difference, demonstrating that the word-final /l/s have a higher degree of velarization than word-final /ɹ/s.

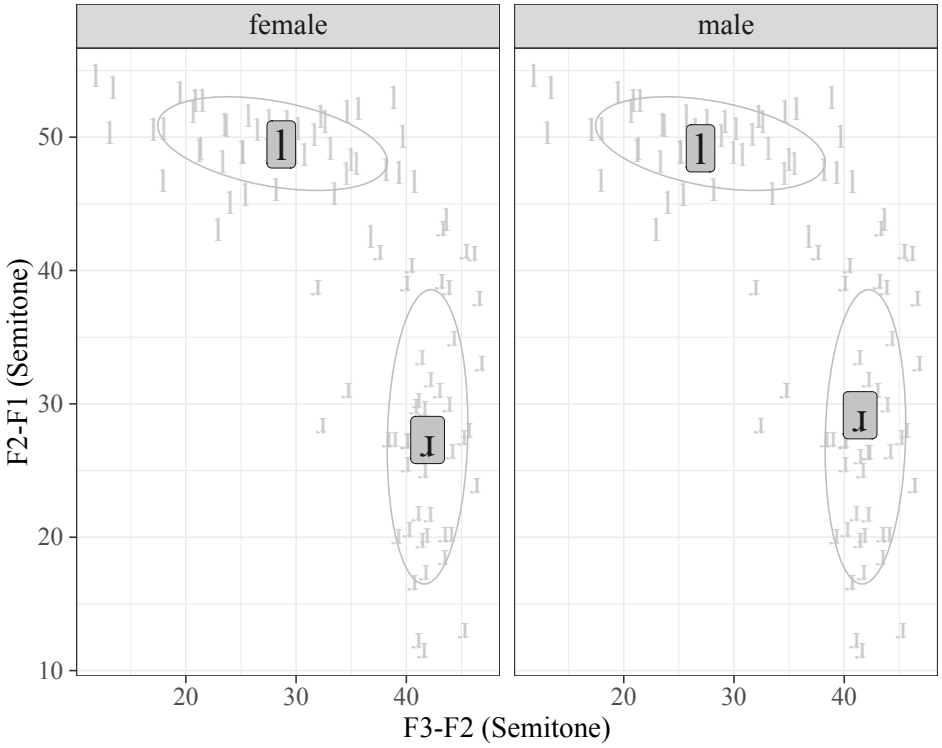

Figure 3.2: Formant Information of L1 Liquids

Formant Information of L2 Liquids

Eight L2 speech samples were selected based on the vetted transcriptions, five of which involved the replacing of English word-final /ɹ/ with a trill /r/; the remaining three were /l/-related variations: two of the three replaced English word-final /l/ with a retroflexed /l̢/, the third replaced /l/ with a flap /ɾ/. Labeling of L2 segments followed the same method as mentioned above. F1, F2, and F3 values were similarly extracted with PRAAT. Figure 3.3 demonstrates the spectral information of both the L1 and L2 segments, while the phonetic symbols with a gray background represent L1 segments and the phonetic symbol without background represent the eight L2 segments.

As Figure 3.3 shows, both of the retroflexed [l̢]s have a relatively smaller F3-F2 difference than English [l]s, showing that the [l̢]s are likely to be more rhotic than their L1 counterparts. The L2 /r/s, on the other hand, have a relatively larger F3-F2 difference than the L1 /ɹ/s, showing that the L2 /r/s

are less rhotic than their L1 counterparts. The eight L2 speech samples were therefore considered as correctly transcribed. They were consequently chosen to represent L2 liquid mismatches.

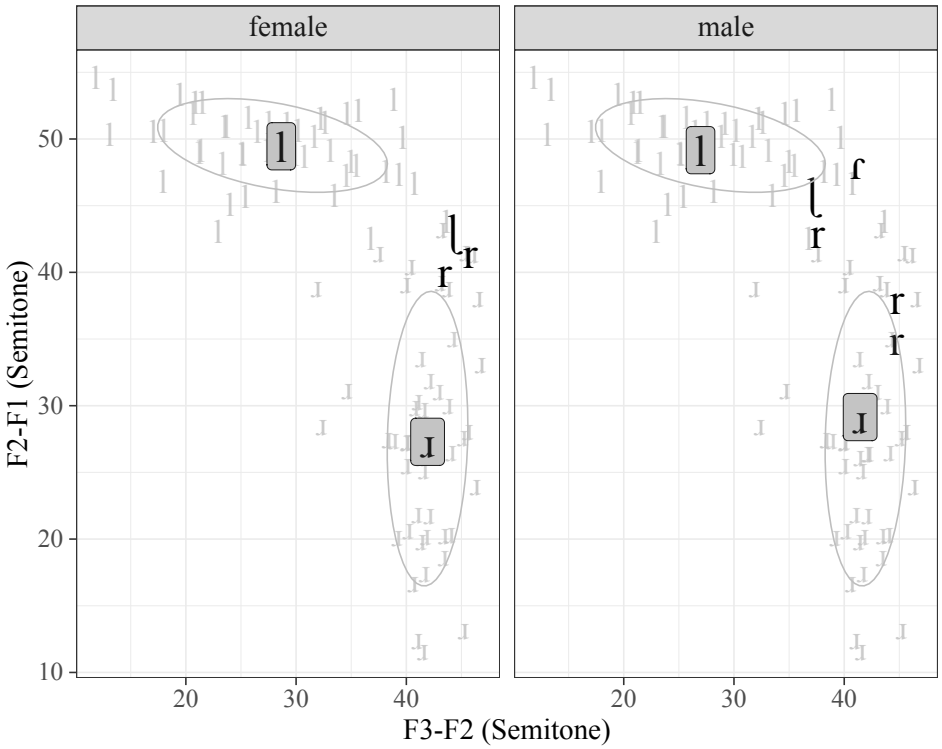

Figure 3.3: Formant Information of L2 Liquids

3.4.5 Vowels

Labeling of vowels was done in PRAAT. The transitioning periods in and out of a vowel were included in the label. Following methods described in De Jong and Wempe (2009), F1 and F2 values were estimated at the location of the energy peak in the amplitude spectrum of a vowel. Female spectral information was extracted at a pitch range from 100 Hz to 500 Hz, with the maximum frequency set at 5500 Hz. Male spectral information was extracted at a pitch range from 75 Hz to 300 Hz, with the maximum frequency set at 5000 Hz. Formant frequency Hertz values were converted to semitones relative to 100 Hz.

Since formant frequencies of a vowel vary considerably depending on the adjacent phonological context, the comparisons between L1 and L2 vowel productions were therefore carried out in each phonological context. The following sections focus mainly on vowel space, a two-dimensional area bounded by lines connecting the F1 and F2 coordinates of a vowel. F1 and F2 values are inversely related to vowel height and vowel frontness. Prosodic elements such as vowel intensity, pitch contour and vowel duration are not discussed.

Context 1: "*ask her*"

According to data in the SAA, the most common L1 production for "*ask*" is [æsk], which was considered by the current study as the L1 target production. L2 speech samples that were transcribed as [æsk] were termed as the match stimuli. L2 speech samples that were not transcribed as [æsk] were the mismatch stimuli. The current study selected five L2 speech samples that contain vowel mismatches. These L2 speech samples were transcribed as [ɐ̟sk həɹ] (i.e., vowel lowering), [ɐ̟sk həɹ] (i.e., vowel raising), [ɑsk həɹ] (i.e., vowel backing), and two incidences of [a:sk həɹ] (i.e., vowel lowering and lengthening).

To evaluate whether these IPA transcriptions are reliable, F1 and F2 values were extracted in the manner described above. Figure 3.4 illustrates the formant frequencies of the 50 L1 and the five L2 productions of /æ/, where the F2 values represent vowel frontness and the F1 values represent vowel height. The /æ/s in gray squares represent mean values of L1 productions. The small gray symbols represent the 50 L1 speakers' productions. The five large emboldened symbols represent the five L2 productions of /æ/.

Figure 3.4 shows that the vetted transcriptions are generally accurate in describing the L2 segments. Discrepancies, however, do exist between the current production analysis and the transcribers' perception. Spectral information indicates that the L1 /æ/s are more front than all the L2 productions. Therefore, the difference between the L1 [æ]s and the L2 [a]s, should lie in the frontness of the vowel, rather than the height. Despite the discrepancies, all the L2 vowels are at least one standard deviation apart from the L1 means (i.e., the shaded area), showing that the L2 vowels are very likely to be different from the L1 target production /æ/. Transcriptions of the five

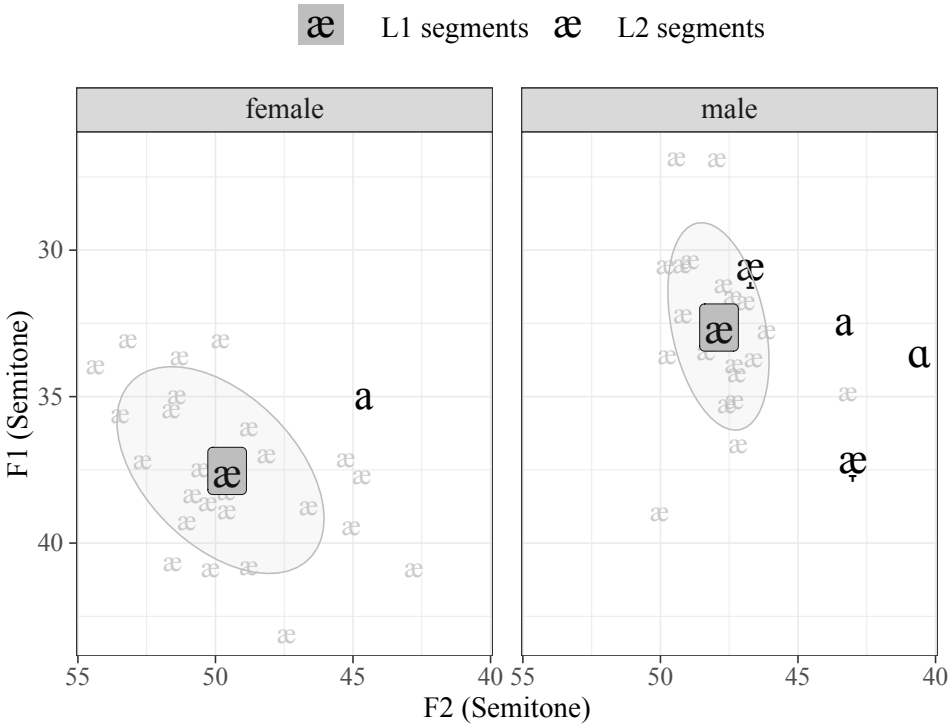

Figure 3.4: Formant Comparison between L1 and L2 Vowels in "*ask her*"

L2 stimuli were accepted by the current study as representations of vowel mismatch. The two L2 [a]s were considered a problem of vowel backing, rather than vowel lowering.

Context 2: "*small plastic*"

In the context "*small plastic*," three types of vowel mismatches were investigated, namely vowel raising, vowel fronting and vowel lowering. Two L2 speech samples were selected to represent the raising of [ɑ] in "*small*." The two L2 vowels for [ɑ] were transcribed as [o] and [ɔ]. The L1 target production for "plastic" was determined as [pʰlæstɪk]. One L2 speech sample was chosen to represent the fronting or the tensing of [ɪ]. The L2 vowel for [ɪ] was transcribed as [i]. Two other L2 speech samples represent the lowering of [æ] in "*plastic*." These two L2 vowels were transcribed as [a] and [a̟]. Figure 3.5 illustrates the formant information of the L1 and L2 segments.

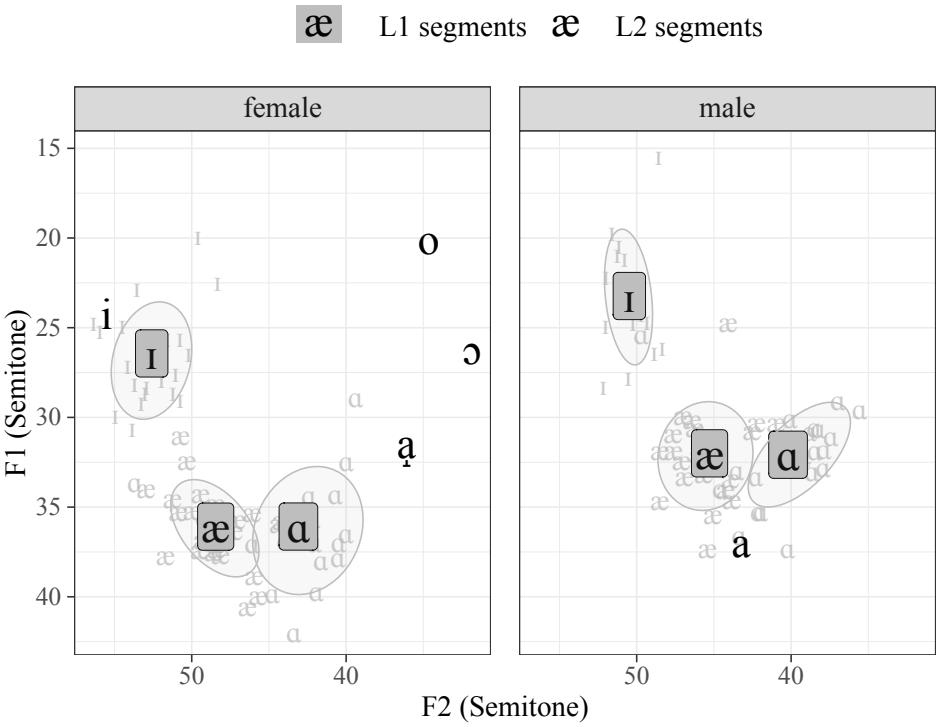

Figure 3.5: Formant Comparison between L1 and L2 Vowels in "*small plastic*"

Figure 3.5 illustrates that the IPA transcriptions for L2 segments are accurate in describing formant information of the L2 segments. These five L2 speech samples were therefore chosen to represent vowel raising, vowel fronting, and vowel lowering in the context of "*small plastic*."

Context 3: "*please call*"

In the context of "*please call*," five L2 speech samples were chosen to represent the raising of /ɑ/. Only the productions of the vowel /ɑ/ in "*call*" were investigated. The five L2 productions of the vowel /i/ in "*please*" were all transcribed as [i], which matches transcriptions of the match stimuli. Therefore, analysis in this section focuses only on the vowel in "*call*". Formant frequencies of L1 and L2 vowels are plotted in Figure 3.6, where the two large emboldened /ɑ/s epresent the L1 mean values. The small gray symbols represent the 50 L1 speakers' productions of /ɑ/. The [ɔ]s and [o]s

represent the five L2 productions of /ɑ/. Figure 3.6 shows that the L2 vowels for /ɑ/ are indeed more back and higher than the L1 norms.

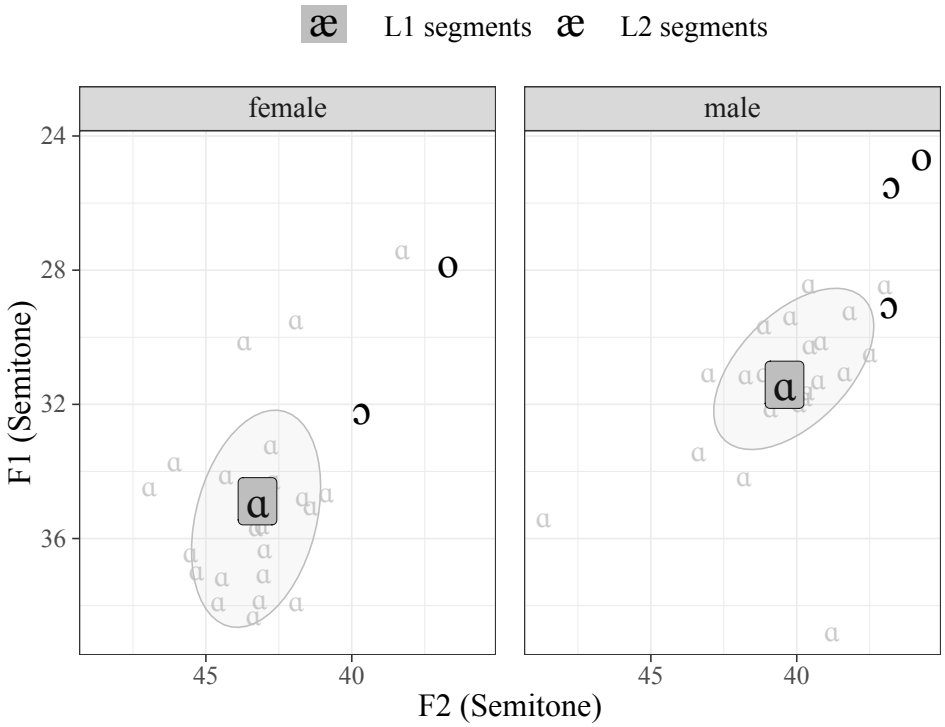

Figure 3.6: Formant Comparison between L1 and L2 Vowels in "*call*"

Context 4: "*five thick*"

Five L2 speech samples were chosen to investigate the vowels in the context of "*five thick*," three of which contain L2 vowels for the lax vowel /ɪ/ in "*thick*," while the remaining two contain L2 vowels for the diphthong /aɪ/ in "*five*." Figure 3.7 demonstrates the vowel space of the L1 and L2 vowels. Four of the five L2 vowels are at least one standard deviation away from the L1 means. The fifth L2 vowel was transcribed as [a], and its vowel space is very close to the L1 mean of /aɪ/.

The SAA classified the L2 [a] as a shortened L2 variation of the L1 [aɪ]s. The difference between [a] and [aɪ] is that [aɪ] contains an off-glide, which requires the raising of the tongue near the end

of the articulation. The English [aɪ] is a falling diphthong, which starts with a segment that carries higher prosodic prominence (i.e., higher pitch and/or loudness) and ends in an off-glide that carries less prominence. As mentioned above, the current study opted to extract F1 and F2 values at the spectral peak of a vowel. The vowel space of [aɪ] illustrated in Figure 3.7, therefore, mainly shows the vowel space of the [a] in [aɪ], which is normally where the spectral peak is located. Therefore, the F2 and F1 values extracted by the current study cannot successfully distinguish [aɪ] from its L2 segment [a].

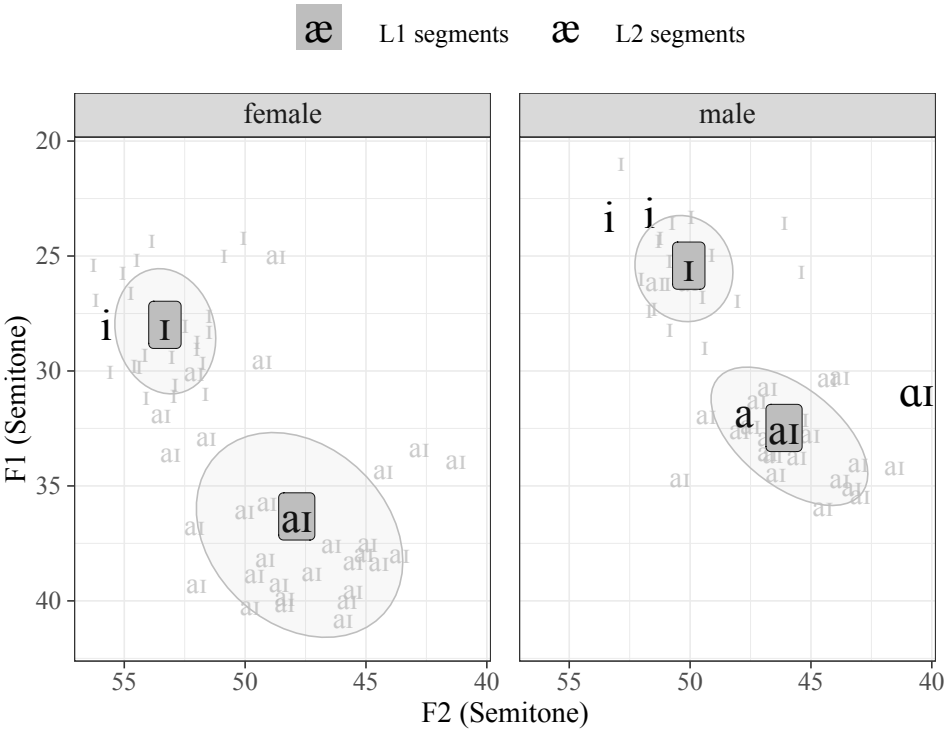

Figure 3.7: Formant Comparison between L1 and L2 Vowels in "*five thick*"

The [a] was produced by a Mandarin speaker. The [ɑɪ] was produced by a German speaker. To measure the dynamic spectral difference between [aɪ] and [a] in the environment of "*five*," dynamic spectral information needs to be investigated. Since [ɪ] is more front and higher than [a], an F1 decrease and an F2 increase is expected toward the end of [aɪ]. F1 and F2 contours of the L2 vowel

[a] would remain relatively flat. Since [ɑɪ] starts with a back vowel [ɑ] and ends in an off-glide, an F2 increase would be expected. However, since [ɑɪ] starts with a back vowel [ɑ], F2 values of [ɑɪ] should be, in general, lower than F2 values of [aɪ].

To evaluate the dynamic spectral difference between the L1 diphthong /aɪ/ and its two L2 counterparts (i.e. [a] and [ɑɪ]), F1 and F2 contours for each L1 [aɪ] were extracted using PRAAT. The contours were then divided into 10 equally spaced time points, from which F1 and F2 values were extracted and converted to semitones relative to 100 Hz. Smoothing Spline ANOVA (SSANOVA) was implemented on the F1 and F2 contours of L1 [aɪ]s using the *gss* package in R (Gu, 2014) with 95% Bayesian confidence intervals.

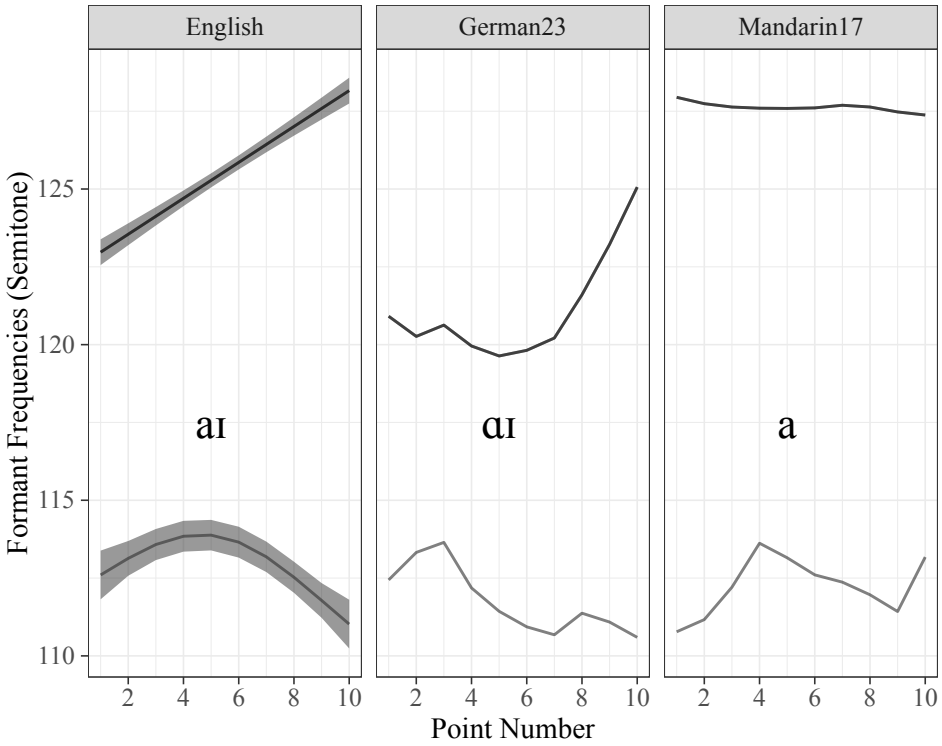

Figure 3.8: Dynamic Formant Comparisons between L1 and L2 Vowels in “Five”

F1 and F2 contours for the Mandarin speaker’s [a] and the German speaker’s [ɑɪ] were similarly extracted. Figure 3.8 shows the F1 and F2 contours of L1 English [aɪ]s and the L2 vowels [ɑɪ] and

[a] [2] Contours on the top represent F2 changes across 10 time points, while the contours on the bottom represent F1 changes across 10 time points. As expected, the L1 [aɪ] exhibits a rising F2 contour, while the L2 vowel [a] shows a relatively flat F2 contour. The L2 vowel [ɑɪ] also exhibits a rising F2 contour. Its general F2 values, however, are lower than the L1 [aɪ]s, showing that L2 [ɑɪ] are indeed more back than [aɪ]s. These results show that the IPA transcriptions for the two L2 vowels have successfully captured the spectral differences between the two L2 vowels and their corresponding L1 productions. These two L2 speech samples were accepted as representations for off-glide deletion and vowel backing in the context of "*five thick*."

Context 5: "*six spoons*"

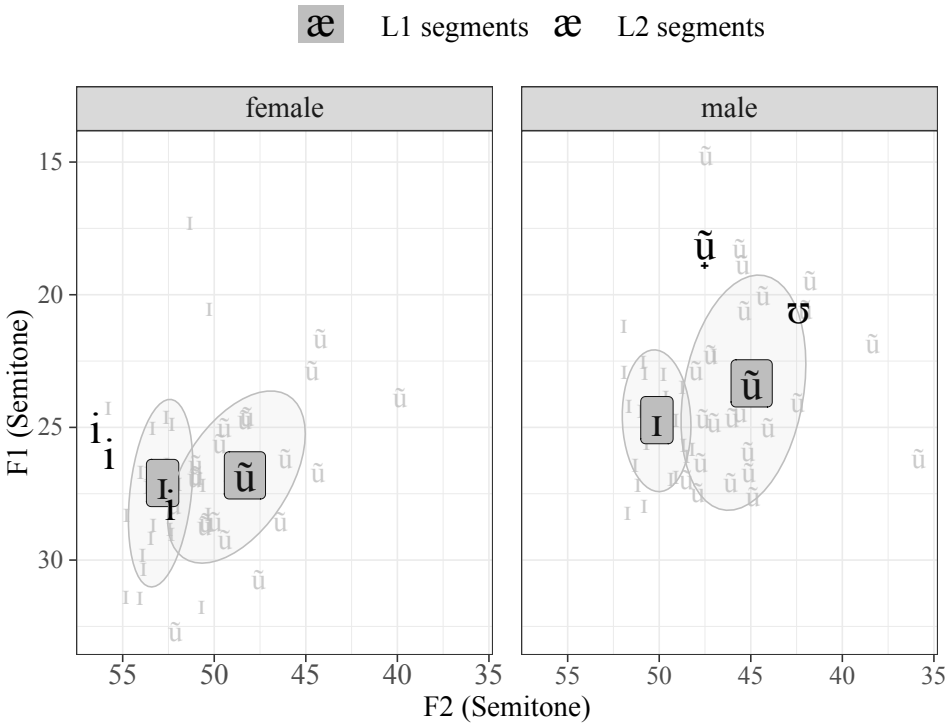

Figure 3.9: Formant Comparison between L1 and L2 Vowels in "*six spoons*"

[2]Note: Both the German and Mandarin speakers are male. The "English" contours represent mean values of 25 male L1 speakers.

Five L2 speech samples were selected for the investigations of [ɪ] and [ũ] in "*six spoons*," three of which represent the tensing or fronting of [ɪ] to [i], two represent the raising or de-nasalization of [ũ]. Vowel spaces of the L1 and L2 vowels are plotted in Figure 3.9. The finding that [ʊ] is higher than [ũ] is not unexpected, because nasalized vowels are often observed to have significantly lower F1 values (Carignan, 2017; Hawkins and Stevens, 1985).

Figure 3.9 also demonstrates that four of the five speech samples contain vowels that are quite different from the native norms. One L2 speech sample that contains a [i] that is spectrally very close to the mean of L1 [ɪ]s. A question remains as to why this specific L2 segment was perceived as a tensed vowel by SAA transcribers. Voice quality was found as a correlate of English tense/lax contrast. The following section measures voice quality of the L2 [i].

Voice Quality of [i] Formant information mainly approximates how the vocal tract modifies the sound made by the vocal folds. The movement of vocal folds generates sounds that could be perceptually "breathier" or "creakier". Both breathy and creaky voice could be produced with a certain amount of vocal fold vibration (i.e., voiced sounds). Compared to creaky voice, breathy voice is produced with a relatively larger space between the vocal fold (i.e., larger glottis opening). The difference in voice quality, such as "breathiness" or "creakiness", is of some relevance to the distinction between English lax and tensed vowels.

Research on English tense and lax vowels has shown that voice quality might be used to perceptually distinguish the English tense-lax contrast in the absence of F1/F2 difference (Di Paolo and Faber, 1990; Lotto, Holt, and Kluender, 1997). More specifically, a high front vowel with more breathy quality is more likely to be perceived as the tense vowel /i/.

To investigate whether voice quality has affected SAA transcribers' perception of tensing, it is necessary to measure the voice quality of the lax vowel [ɪ] in "*six*." A reliable acoustic parameter for voice quality in many languages is spectral tilt, which measures the degree of amplitude change as frequency increases. Previous research has identified several acoustic parameters as measurements for voice quality (Garellek, 2019). The current study opted to quantify spectral tilt by comparing the amplitude of the first harmonics (H1) to the second harmonics (H2), which is the most commonly

used and typically most reliable measurement for voice quality.[3] Breathy voice usually correlates with stronger amplitude on the lower harmonics and weaker amplitude on higher ones. Creaky voice is the opposite. Voice quality, therefore, is often measured by subtracting H2 from H1.

L1 and L2 speech samples of "*six spoons*" were analyzed using the VoiceSauce package (Shue et al., 2011) in MatLab R2017b. The frequencies of harmonics and fundamental frequencies were estimated by the STRAIGHT algorithm (Kawahara et al., 2009). The Snack Sound Toolkit (Sjölander, 2004) was used to locate the formants.

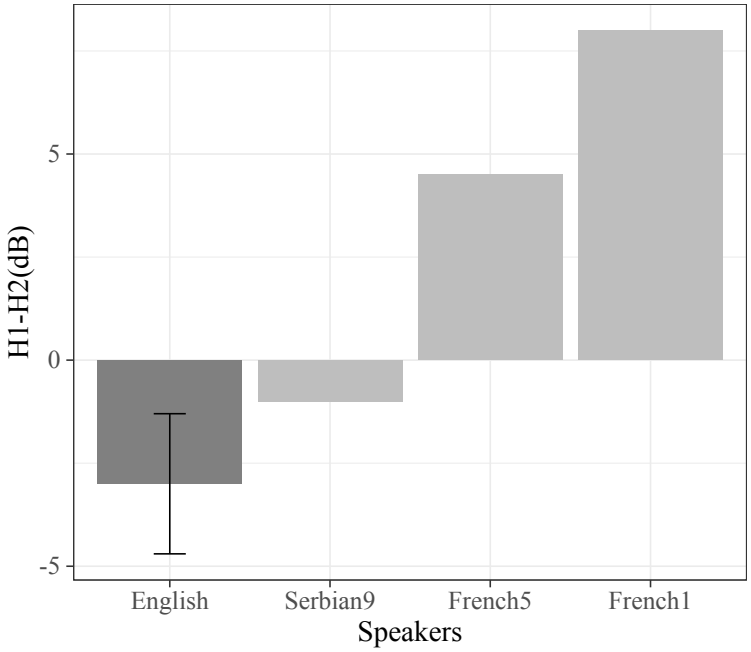

Figure 3.10: Voice Quality Comparisons between L1 and L2 Productions of /ɪ/

Figure 3.10 shows the average H1-H2 values of L1 and L2 segments. The three L2 speech samples involving the raising of [ɪ] in "*six*" were all produced by females. Therefore, the comparison below illustrates only the difference between the three female L2 productions and all female L1 productions. As the figure shows, all three L2 speakers' pronunciation had a relatively higher degree of

[3]Several studies reported that H1-H2 might not be reliable in nasal environments. See Garellek (2019) for a more thorough review

breathiness (i.e., higher H1-H2 values) than the L1 speaker norm. The Serbian speaker's production is less breathy than the other two L2 speakers; it is nevertheless breathier than the native mean. The fact that the SAA transcribers perceived high vowels with breathy quality as [i] is consistent with previous findings on the effect of voice quality on vowel height perception (Lotto, Holt, and Kluender, 1997). The IPA transcriptions for these L2 speech samples were therefore accepted by the current study to represent vowel tensing.

3.4.6 Summary of Segmental Analysis

The current study selected four types of L2 speech samples. Stimuli selection was based on IPA transcriptions available in the SAA. The most common L1 productions in the SAA were considered the L1 target productions. L2 speech samples that were transcribed the same as their L1 target productions were termed as the match stimuli. L2 speech samples whose IPA transcriptions differ from transcriptions of their L1 target productions were termed as the mismatch stimuli. The sections above described two types of mismatch stimuli. The first type consists of 25 L2 speech samples that were termed consonant mismatch. IPA transcriptions for these L2 speech samples differ from their L1 target productions by only one consonant. The second type consists of 25 L2 speech samples that were termed vowel mismatch. IPA transcriptions for these L2 speech samples differ from their L1 target productions by only one vowel.

Acoustic analysis was conducted to examine the reliability of the IPA transcriptions. As shown in the analysis above, acoustic differences between the L2 speech samples and their corresponding L1 speech samples could be captured by IPA transcriptions of the L2 speech samples. The current study, therefore, considered these IPA transcriptions reliable. Audio files of these L2 speech samples were thus used by the current study as stimuli in two perception studies (Chapters 4 and 5).

The third type of mismatch was termed syllable mismatch, because it concerns syllable structure differences between an L2 speech sample and its L1 target production. While consonant or vowel changes preserve the original syllable structure, segment deletion or segment epenthesis changes the original structure. The next section describes the selection of the 25 speech samples with syllable mismatches.

3.4.7 Syllable Structures

Structural variations investigated by the current study involve vowel epenthesis and consonant deletion. Co-articulatory properties and/or L1 phonotactics could, to some degree, generate a perceptual illusion of segment insertion, often termed as the "ghost segments" or "illusory segments" (Dupoux et al., 1999). To select stimuli with potential structural problems, L2 audio samples were inspected in PRAAT to rule out possible cases of illusory perception. The following section uses two examples to illustrate how the inspections were carried out. As mentioned previously, the most common L1 productions in the SAA were considered representatives of L1 target productions. For example, the majority of the 100 surveyed L1 speakers of American English pronounced the word "*ask*" in phrase "*ask her*" as [æsk], while only one of the 100 L1 speakers pronounced the "*ask*" as [æs]. [æsk] was therefore considered the L1 target production. [æs] was considered as a production with syllable mismatch. In other words, productions with mismatches are uncommon in L1 speech, but they are not necessarily unique to L2 speakers.

To select L2 speech samples whose syllable structure differ from L1 target productions, the current study selected 25 L2 speech samples from the SAA based on their respective IPA transcriptions. To further inspect the reliability of the IPA transcriptions, acoustic information of the L2 speech samples were examined in PRAAT. If spectral information of a segment was missing from an L2 speech sample but existed in L1 target productions, then the L2 speech sample was considered as a stimulus with syllable mismatch. More specifically, it was considered a stimulus with segment deletion. Alternatively, if spectral information of a segment existed in an L2 speech sample but was missing from L1 target productions, then the L2 speech sample was considered as a stimulus with segment insertion.

Segment Deletion

Figure 3.11 shows the spectrogram of an L1 production and an L2 production of "*ask her*", where the dotted lines represent pitch contours and the solid lines represent intensity contours. The graph on the left illustrates an L1 production, which shows visible stop closure and burst after /s/. These characteristics are absent from the L2 production on the right, indicating that coda /k/ was dropped

by the L2 speaker.

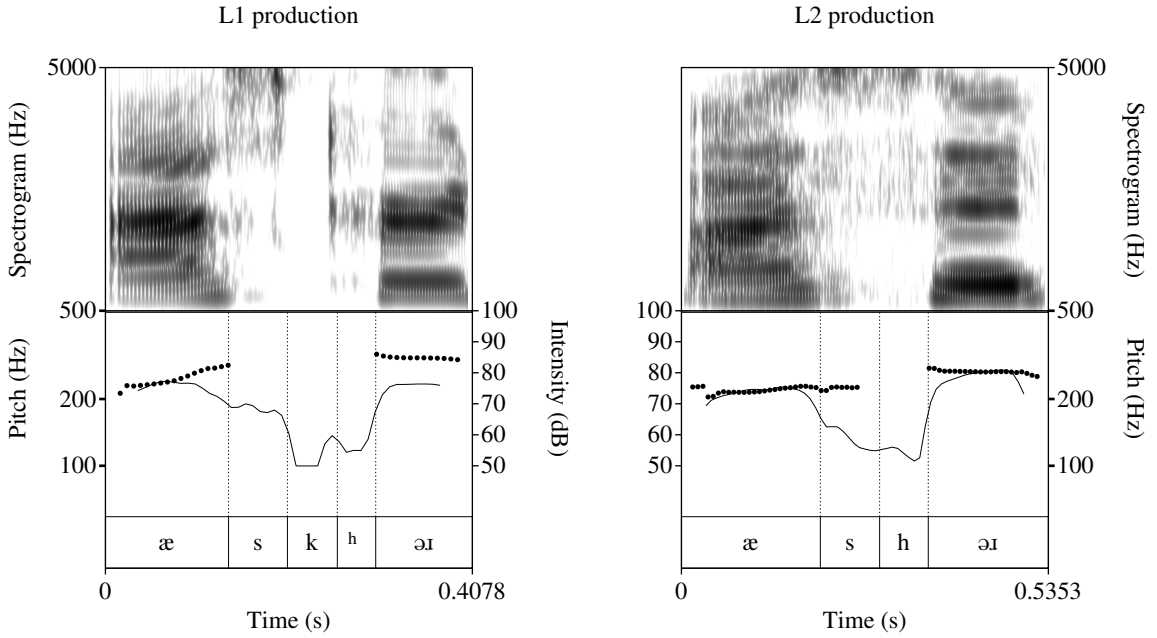

Figure 3.11: /k/-Deletion in "*ask her*"

Segment Insertion

Eight stimuli were included in the current study to represent three types of segment insertions. Two stimuli involved prothesis of s-clusters (i.e., [sp]→[əsp] in "*six spoons*"); three stimuli involved anaptyxis of /pl/-clusters (i.e., /pl/ to /pəl/ in "*please call*"); two stimuli represent paragoge at the end of "*ask*" (i.e., [æsk]→[æskə]), and one stimulus represents paragoge at the end of "five" (i.e. [faɪv] to [faɪvə]). The SAA transcribers marked discourse fillers and epenthetic vowels differently. A space was added between discourse fillers and their adjacent segments (e.g., [æsk ə həɪ]). No space was added between epenthetic vowels and the segment they epenthesize to (e.g., [æskə həɪ]). Speech samples with discourse fillers were not selected. The eight stimuli all contain epenthetic vowels as indicated by their respective IPA transcriptions.

Figure 3.12 illustrates a case of paragoge. The speech sample was produced by a Korean speaker

who inserted a /ə/-like vocoid at the end of the word "*ask*." The epenthesized vocoid was transcribed by the SAA transcribers as a [ə].

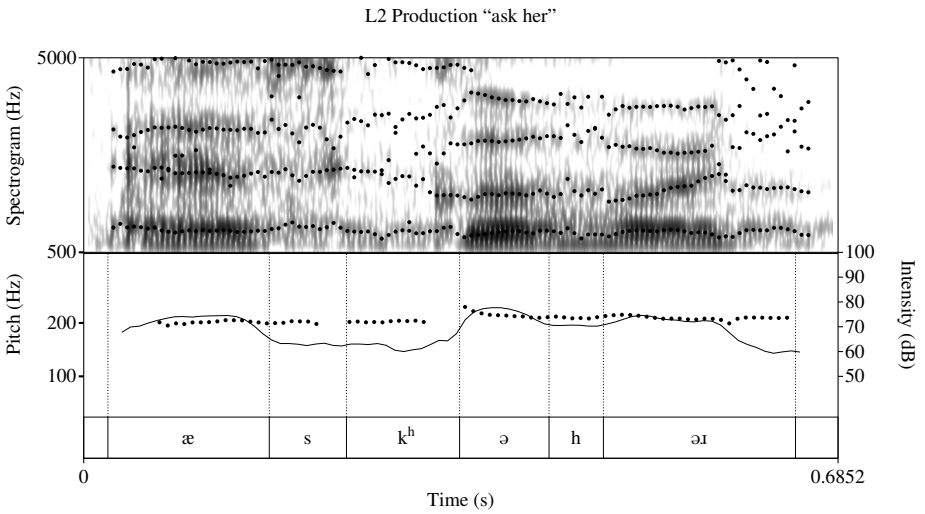

Figure 3.12: Paragoge after "*ask*"

Spectral inspection carried out by the current study has successfully identified the epenthesized vocoid because it shows clear formant structures, carries pitch, and contains an intensity peak. These characteristics were utilized in the inspection of all other cases of segment insertion. Of the eight cases inspected, seven of them satisfy the aforementioned criteria. The prothetic vocoid of /sp/ does not carry pitch, yet it contains clear formant structures and an intensity peak. One could argue against defining such vocoid, and indeed all the other seven vocoids, as epenthetic vowels. As previous research often shows, epenthetic vowels, transitional vowels and extended sonorants are sometimes difficult to distinguish, and definitions of epenthetic vowels vary from language to language (Gouskova and Hall, 2009; Hall, 2003, 2011).

The current study took an impressionistic approach with regard to vowel epenthesis. As long as the epenthesized segment has clear formant structure and was transcribed with a [ə] or [ɪ], it is considered an epenthetic vocoid. In summary, IPA transcriptions of the 25 L2 speech samples were verified via the process discussed above. The 25 L2 speech samples were selected as stimuli with syllable mismatches.

3.5 Summary

The current study selected 100 L2 speech samples as stimuli in two perception studies. The L2 speech samples were chosen based primarily on their IPA transcriptions in the SAA. To observe the potential effects of phonological context on accentedness perception, five phonological contexts from the so-called "Stella" passage were chosen. Each context was represented by 20 L2 speech samples, yielding 100 L2 speech samples in total. To determine the L1 target productions for the five contexts, IPA transcriptions of 100 L1 speakers of American English from the SAA were surveyed to find the most common productions (e.g., [pʰliz kʰɑl] for "*please call*"). L2 speech samples that were transcribed the same as their L1 target productions were termed as the match stimuli. L2 speech samples that were not transcribed as the same as their L1 target productions were termed as the mismatch stimuli.

Among the 20 L2 speech samples for each of the five contexts, five speech samples were the match stimuli (e.g., [pʰliz kʰɑl] for "*please call*"). The rest 15 L2 speech samples were the mismatch stimuli. Among the 15 mismatch stimuli, five differed from their L1 target production by only one consonant (e.g., [pʰlis kʰɑl] for "*please call*"), five differed from their L1target production by only one vowel (e.g., [pʰliz kʰol] for "*please call*"). Another five L2 speech samples, did not differ from their L1 target production segmentally, but contained either one more or one less segment than their L1 target productions (e.g., [pʰəliz kʰɑl] or [pʰliz kʰɑ] for "*please call*"). These three types of stimuli were termed consonant mismatch, vowel mismatch, and syllable mismatch respectively.

To verify the reliability of the IPA transcriptions for the 100 L2 stimuli, acoustic analysis was performed to compare L2 speech samples with their L1 counterparts. Speech samples from 50 L1 speakers of American English were extracted from the SAA for the analysis of native speaker pronunciation norms, which were approximated by the mean L1 values of relevant benchmark acoustic measurements (e.g., mean L1 VOT duration, mean L1 COG, mean L1 F1/F2/F3 values etc.). Results showed that acoustic differences between the L2 stimuli and native speaker norms were successfully captured by the IPA transcriptions. The current study therefore concluded that these IPA transcriptions for the 100 L2 stimuli are reliable. These stimuli were then used in two perception studies to elicit accentedness judgment from L1 listeners of American English.

Chapter 4: Experiment 1

4.1 Introduction

Experiment 1 of the current study investigates the accentedness of various phonetic patterns in non-native (L2) English speech. The current study considers the most common native (L1) English productions observed in one hundred L1 speakers of American English as the L1 target productions (e.g., [æsk] for "*ask*"). L2 speech productions that differ from the L1 target productions were considered mismatches. The mismatch stimuli, therefore, represent patterns in L2 speech that do not match the most common L1 productions.

Eleven types of consonant mismatches, 5 types of vowel mismatches and 2 types of syllable structure mismatches were assembled to enable a detailed comparison between different types of speech patterns in L2 speech. The stimuli were chosen based on their phonetic transcriptions (IPA transcriptions), which were vetted by at least 3 professional transcribers and were further examined by acoustic analysis conducted by the current study, as described in Chapter 3.

L1 American English raters were recruited from the Amazon Mechanical Turk (MTurk) platform to provide accentedness judgments on the stimuli. The results provide direct comparisons between the accentendess of different types of consonant, vowel and syllable structural patterns in L2 speech.

4.2 Stimuli

Stimuli for the current study were selected based on their respective phonetic transcriptions, which were verified by acoustic analysis conducted as part of the current study (Chapters 3). An example of the four types of stimuli were introduced in Chapter 3 and are re-listed here in Table 4.1, where the "Contexts" column specifies the five phonological contexts. There are 20 stimuli for each of the 5 contexts, yielding 100 stimuli in total.

Table 4.1: Types of Stimuli

Contexts	Match	Consonant Mismatch	Vowel Mismatch	Syllable Mismatch
please call	[pʰliz kʰɑl]	[**p**liz kʰɑl]	[pʰliz kʰ**o**l]	[pʰ**ə**liz kʰɑl]
ask her	[æsk (h)əɹ]	[æsk hə**r**]	[**ɑ**sk həɹ]	[æs_ həɹ]
six spoons	[sɪks spunz]	[sɪks spun**ʃ**]	[s**i**ks spunz]	[sɪks **ə**spunz]
five thick	[faɪv θɪk]	[faɪv **t**ɪk]	[f**a**v θɪk]	[faɪv**ə** θɪk]
small plastic	[smɑl pʰlæstɪk]	[smɑ**l̩** pʰlæstɪk]	[smɑl pʰlæst**i**k]	[smɑl pʰlæs_ɪk]

IPA transcriptions of the match stimuli are the same as IPA transcriptions of their L1 target productions (e.g. [θɪk] for "*thick*"), meaning that the IPA transcriptions of the match stimuli match IPA transcriptions of their L1 target productions. L2 speech samples that do not match their L1 target productions were termed the mismatch stimuli. The mismatch stimuli were further divided into three groups based on three types of mismatches, namely, stimuli with consonant mismatches (e.g., [tɪk] for "*thick*"), stimuli with vowel mismatches (e.g., [θik] for "*thick*"), and stimuli with syllable structure mismatches (e.g., [faɪvə] for "*five*"). Experiment 1 therefore investigates four types of stimuli (i.e., the match stimuli and the three types of mismatch stimuli).

4.3 Procedure

Experiment 1 recruited L1 listeners of American English to provide accentedness ratings of the 100 stimuli. Participants (i.e., raters) heard each of the 100 audio stimuli and were then asked to judge the degree of the foreign accent exhibited in the stimulus on a 9-point Likert-like scale (Figure 4.1). Following the practice of similar studies (e.g., McCullough, 2013), only the endpoints of the scale were marked.

The raters heard the 100 audio stimuli without knowing the intended meaning of the stimuli. A rating of one means the stimulus has no foreign accent at all. A rating of nine means the stimulus has a very strong foreign accent.

The interface of the experiment provided a button and a 9-point rating scale. Raters were instructed to wear headphones or earbuds to listen to the stimuli[1]. A stimulus was played once the rater hit the button, after which the rating scale would appear. Raters provided their accentedness judgment by choosing a number from one to nine on the rating scale, and then moved on to the next trial. There was no time limit for each trial. The maximum time allowed for completing the entire experiment was 30 minutes. Figure 4.1 illustrates the interface of the experiment.

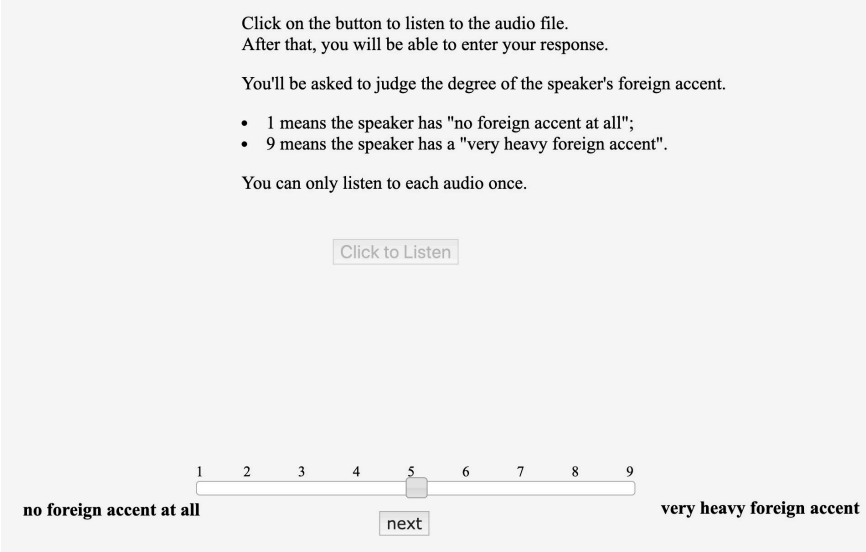

Figure 4.1: Interface of Experiment 1

To reduce the order-effect, a block randomization technique was implemented. The 100 stimuli were divided into five blocks, each of which contained one token per type per context, yielding 20 stimuli per block (five contexts × four types). The order of blocks and the stimuli in each block were randomized for each participant via JavaScript using the Fisher-Yates Shuffle algorithm (Fisher and Yates, 1963).

Raters of Experiment 1 were not required to identify or locate the specific type of mismatch in each stimulus, because the mismatch had already been determined by the vetted transcriptions.

[1] Since the experiment was conducted online, we cannot ensure that all the raters wore headphones or earbuds during the experiment. This could be a potential methodological limitation. Readers could consult Woods et al. (2017) for a psychophysical test that helps to determine whether online experiment participants are wearing headphones.

There were 100 trials in total. At the end of the experiment, the raters were asked to take a demographics survey, which collected information on the raters' age, gender, L1/L2 status, occupation, current residence and birthplace. Raters on average spent 12.3 minutes (SD=3.2 minutes) on the experiment. Raters were compensated $0.50 upon completion of the experiment. The experiment was programmed with HTML, CSS and JavaScript.

Previous studies often provide a training session to familiarize participants with the range of accents in the experiment (Major, 1986; Munro, Derwing, and Flege, 1999). However, the conundrum is that there is no way to obtain the full range of accents without testing the raters first. Experimenters could subjectively select a few representative stimuli for the training session, but experimenters' own biases could be introduced in the process. Experiment 1 therefore opted to omit the training session. The Result section of this chapter discusses how the absence of training affects raters' accentedness judgments.

4.4 Raters

Participants (i.e., raters) were 110 adult L1 American English speakers recruited via Amazon Mechanical Turk (MTurk), a web-application that allows researchers to conduct survey-based experiments. Previous literature has shown that results of behavioral experiments conducted on MTurk are comparable to results of similar experiments conducted in lab settings (Enochson and Culbertson, 2015; Sprouse, 2010). Difallah, Filatova, and Ipeirotis (2018) recently showed that there are about 2,000 participants active on MTurk at any given time. 51% of them are female, 49% of them are male. About 75% of the participants are from the United States. Indian participants represent 16% of the population. The rest are from Canada, Great Britain, the Philippines and Germany.

Since Experiment 1 aims to investigate accentedness judgments of American English listeners, the experiment was made accessible only to people with a U.S. IP address. To increase the reliability of responses, the experiment required participants to have an approval rating of at least 95%. That is, at least 95% of each participant's previous work on MTurk has met the requirements of the people who assigned the work. All of the participants reported their birth location and residence as being in the United States. All of them reported that that they are L1 speakers of English. We therefore

assumed that the participants are L1 speakers of American English. Two of the participants reported having speech or hearing related disorders. Responses from these two participants were thus removed, yielding 108 participants in total, among which 61 were female, 45 were male, and two did not report their gender. The age of the 108 participants ranged from 20 to 66. The mean age was 33.50 (SD=12.51) (See rater demographics in Appendix B).

4.5 Control for Prosody

The current study focuses specifically on segmental and syllable structure information. However, prosodic information has often been shown to affect foreign accent perception (e.g., Magen, 1998; Munro and Derwing, 1995). It is therefore important for the current study to control for prosodic information of the selected stimuli. The following section first reviews the methods used by previous studies and then describes the method implemented by the current study.

Two types of methods were often used in previous research to account for prosodic information. One method, usually termed *prosody cloning* or *prosody transplantation*, involves the superimposing of prosody of one utterance onto another (Mareüil and Vieru-Dimulescu, 2006; Yoon, 2007). At least two utterances are required with this method. Usually, highly controlled read speech samples are used, so that the two utterances contain exactly the same number of segments. The duration, fundamental frequency (F0), or intensity of segments in one utterance (the "donor") can be automatically superimposed on the other utterance (the "recipient") via the PSOLA algorithm (Moulines and Charpentier, 1990). In this way, the recipient sample may have similar or even identical prosodic characteristics as the donor.

The drawback of this method is that the number of segments in the donor and the recipient should be the same, which might cause problems for imposing L1 prosody on L2 speech with epenthesis or segment deletion. Moreover, acoustic manipulations of prosody will probably alter the acoustic characteristics in segmental dimensions, which might artificially increase or decrease accentedness of the original speech. Therefore, it might not be ideal to implement this method in studies that investigate segmental characteristics of foreign accent.

An alternative method measures the prosodic differences without acoustic manipulation. This

method implements the Dynamic Time Warping (DTW) alignment algorithm to account for the prosodic difference between two utterances (Adami et al., 2003; Rilliard, Allauzen, and Mareüil, 2011; Sharpe, Fogerty, and Den Ouden, 2015). DTW is a non-linear algorithm that looks for the dissimilarity between two temporal sequences of data and calculates the costs to align one with the other. It generates a DTW score that represents the dissimilarity between two sets of data. The greater the DTW score, the more dissimilar the two sets of data are.

The advantages of the DTW method are that (1) it does not involve any acoustic manipulation of the original speech files; (2) it does not require that the two utterances compared contain the same number of segments, which make it suitable for comparing L1 and L2 speech samples that differ in syllable structure; and (3) the calculation not only compares F0 and/or intensity values at different time points, but also takes into account the durational difference between two utterances, where larger differences will lead to a greater DTW score. In other words, the DTW score might, to some degree, represent the global prosodic differences of two utterances. The DTW method could thus be more suitable for the current study.

In the current study, the DTW algorithm took F0 values of L1 speech samples as the reference and F0 values of an L2 speech sample as the input. The algorithm generated DTW scores, which represent the intonational and durational dissimilarities between the L1 and L2 speech samples. L1 productions were first analyzed to find L1 F0 contours. Male and female speech samples were separately processed. For each gender, speech samples were grouped by the five contexts. F0 contours of the speech samples were extracted using the PRAAT auto-correlation algorithm (Boersma and Weenink, 2015). The contours were then divided into 300 equally spaced time points for every second (i.e., 3.33 ms per point), from which an F0 value was extracted. Following Morrill (2015), gaps in the contour were interpolated from the points on either side of the gap. Artifacts were removed by smoothing with a bandwidth of 5 Hz. F0 values were then converted to semitones relative to 100 Hz. Pitch range for female speech samples was set between 100 Hz and 500 Hz. Pitch range for male speech samples was set between 75 Hz and 300 Hz. To allow for cross-speaker comparison, the semitones were then normalized for each speaker.

F0 values of the 100 L2 stimuli were similarly extracted and normalized. The DTW algorithm

was then implemented in R with the *dtw* package (Giorgino, 2009) to calculate the degree of prosodic dissimilarity between L1 and L2 speech samples. For every L2 stimulus, 25 DTW scores were generated to represent the warping cost between the L2 stimulus and each of its corresponding 25 L1 speech samples. The mean DTW scores for each L2 stimulus were then calculated to account for prosodic information of the stimulus. The mean DTW scores for all of the 100 L2 stimuli were recorded for further statistical analysis.

4.6 Results

4.6.1 Experiment 1 Hypotheses

Based on previous research on speech perception and lexical identification, Experiment 1 hypothesizes that stimuli with consonant mismatches are perceptually more foreign-accented than stimuli with vowel mismatches, because consonants are more susceptible to categorical perception than vowels and are more important in lexical identification. Based on previous empirical findings on L2 speech perception, Experiment 1 predicts that segment epenthesis is perceptually more foreign-accented than segment deletion, because segment deletion is sometimes allowed in L1 speech, while segment epenthesis rarely occurs. Experiment 1 also predicts that stimuli with mismatches should be generally more foreign-accented than stimuli without mismatches (i.e., the match stimuli).

4.6.2 Segmental and Structural Mismatches

The mean ratings across all four stimulus types (each stimulus was rated on a scale from 1 to 9) was 4.81 (SD=2.21). The larger the number, the more foreign-accented a stimulus was judged. As expected, raters assigned higher ratings for stimuli with mismatches (M=5.10, SD=2.15) than for stimuli without mismatches (M=3.94, SD=2.16). Ratings of stimuli with consonant mismatches (M=5.66, SD=2.04) were on average higher than ratings of stimuli with syllable structural mismatches (M=4.96, SD=2.17), which were on average higher than ratings of stimuli with vowel mismatches (M=4.69, SD=2.13). Figure 4.2 demonstrates the mean ratings of the four types of stimuli, where the error bars represent the 95% confidence intervals.

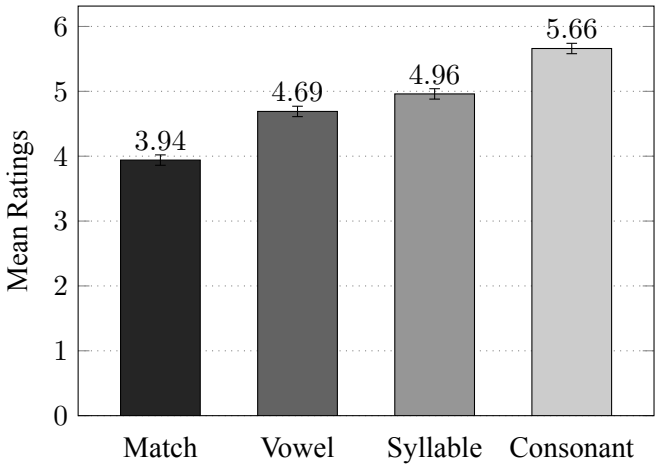

Figure 4.2: Mean Ratings by Type of Stimuli on the Scale from 1 to 9

Linear mixed-effects models were employed with the *lme4* package in R (Bates et al., 2014) to investigate the possible influence of the segmental and syllable mismatches on foreign accent ratings. In the full model, the dependent variable was the ratings. Type of stimuli (consonant mismatch vs. vowel mismatches vs. syllable mismatch vs. match) was Helmert contrast-coded to examine whether the four types of stimuli affect accentedness rating differently. Three stimuli contrasts were created. The first contrast compared consonant mismatch with syllable mismatch; the second contrast compared vowel mismatch with consonant and syllable mismatch; and the third contrast compared the three types of mismatch stimuli with the match stimuli. To investigate accentedness ratings across the 100 trials, trial numbers were also included as a fixed effect. To control for prosody, the mean DTW score of each stimulus was included as a fixed effect. The interactions among the three stimuli contrasts, the trial number and the DTW scores were also included as fixed effects. Raters were included as a random effect with the "type of stimuli" as its random slope. Stimuli were included as another random effect.

The effects of fixed effect factors were investigated by comparing models using the likelihood ratio test. The full model was as described above. The likelihood ratio test is a test of the goodness-of-fit between two models. It produces a chi-square and a p-value to indicate whether one model

fits a particular set of data significantly better than the other model. To investigate the contribution of a particular fixed effect to model fit, the fixed effect needs to be removed from the full model to construct a new model. The new model, therefore, differs from the full model only by the exclusion of one single parameter (i.e., the particular fixed effect factor). If the full model indeed fits the dataset significantly better than the new model, then one could conclude that the exclusion of the specific fixed effect factor significantly changed the model fit, which implies that the fixed effect factor in question is statistically significant.

To investigate the effect of DTW scores, the fixed effect factor DTW cores was removed from the full model to construct a new model. The new model was compared to the full model using the likelihood ratio test. The result showed that the full model did not fit the dataset significantly better than the new model ($\chi 2 = 2.87$, $p = .09$), indicating that the contribution of DTW scores to model fit was not significant. The same method was applied to the investigation on other fixed effects. The results show that none of the interactions involving DTW scores achieved significant contribution to model fit, suggesting that the intonation of the stimuli might not be a major factor affecting accentedness ratings.

The contrast between consonant and syllable mismatches significantly contributed to model fit ($\chi 2 = 6.35$, $p < .05$), showing that stimuli with consonant mismatches were rated as being more accented than stimuli with syllable mismatches. The second contrast, which compares vowel mismatches with consonant and syllable mismatches, contributed significantly to model fit ($\chi 2 = 6.95$, $p < .01$), showing that stimuli with consonant mismatches and syllable mismatches were rated as being more accented than stimuli with vowel mismatches. The third contrast, which compares the match stimuli with the three types of mismatch stimuli, also contributed significantly to model fit ($\chi 2 = 13.32$, $p < .001$), showing that stimuli with segmental and structural mismatches were perceived as being more accented than the match stimuli.

These results suggest that all of the three types of mismatches contributed to the perception of foreign accent. However, stimuli with consonant mismatches were perceived as being more accented than the other two. Among the three types of mismatches, stimuli with vowel mismatches were perceived to be the least accented. The DTW scores, which represent prosodic differences between

L1 and L2 productions, positively correlated with accentedness ratings. The contribution of the DTW scores to model fit was only marginally significant, suggesting that prosodic influence might not be a major factor that affected accentedness ratings.

4.6.3 Ratings across Time

The comparison of models also revealed a significant effect of trial numbers on accentedness ratings (β=0.03, $\chi 2$ = 46.80, $p < .001$), suggesting that ratings were not consistent across the 100 trials. That is, the same stimuli might receive different ratings depending on when it occurred during the experiment. More specifically, the later a stimulus occurred in the experiment, the more accented the stimulus was rated. The interaction between trial number and the third contrast contributed significantly to model fit (β=0.04, $\chi 2$ = 9.87, $p < .01$), showing that stimuli with segmental and structural mismatches became more accented than the match stimuli as the experiment progressed. Figure 4.3 demonstrates the accentedness rating across time, where the ribbons represent the 95% confidence intervals. The accentedness ratings increased as the experiment progressed.

Figure 4.3, in general, confirms the findings in the previous section. That is, stimuli with segmental and structural mismatches were rated as more accented than the match stimuli, and consonant mismatches had higher ratings than the other two types of mismatches. The ratings for all stimuli gradually decreased during the first 20 to 25 trials, and then began to increase.

4.6.4 Individual Mismatches

The analysis above showed that consonant mismatches were judged to be more accented in general. It might be too hasty to draw the conclusion that all consonant mismatches are more accented than the other two types of mismatch stimuli. As mentioned in Chapter 3, stimuli of the current study included 11 types of consonant mismatches, 5 types of vowel mismatches, and 2 types of syllable structure mismatches. The phonological context for each mismatch could have had an effect on whether the mismatch was rated as foreign accented. It is, therefore, important to further examine the possible effect of phonological context on accentedness perception.

Recall that the mismatches occurred in five contexts, namely "*ask her*," "*please call*," "*six*

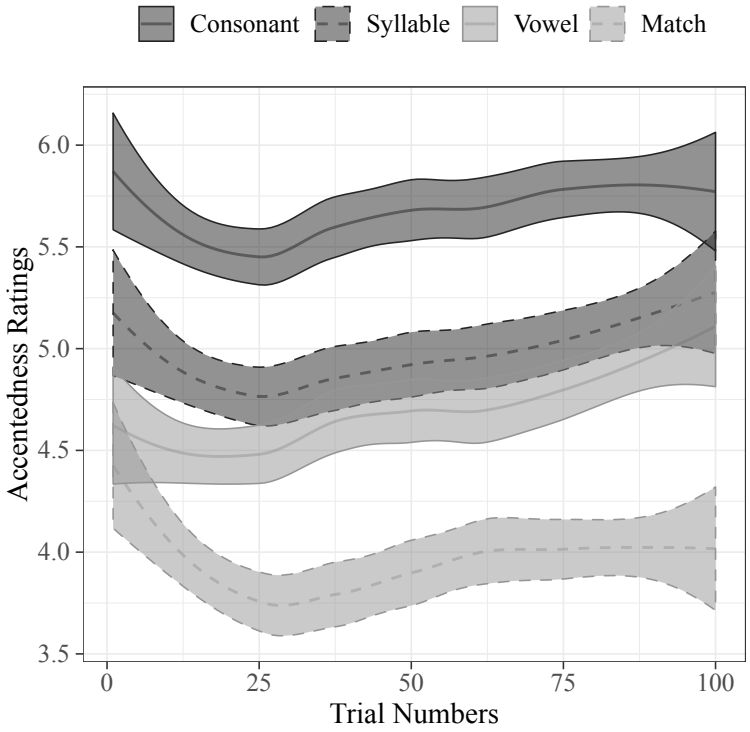

Figure 4.3: Ratings across Time

spoons," "*five thick*," and "*small plastic*." The following session uses context "*ask her*" as an example to discuss the statistical method. Since statistical methods used for the other four contexts were the same as the one used for "*ask her*," the details for these four contexts are omitted.

Context 1: "*Ask her*"

For "*ask her*," there were seven types of mismatch as listed on the y-axis of Figure 4.4. In general, syllable mismatches such as vowel epenthesis (i.e., [æsk]→[æskə]) and /ɹ/-trilling (i.e., [ɹ]→[r]) received the highest ratings. Coda-deletion (i.e., [æsk]→[æs]) received a relatively higher mean rating. Arithmetic mean ratings of the seven types of mismatches and the match stimuli are presented in Figure 4.4, where the error bars represent the 95% confidence intervals.

Mixed-effects linear regression models were constructed to investigate the effect of individual

mismatch on accentedness. The ratings were the dependent variable. The eight types of stimuli and trial number were entered as fixed effects. Raters were entered as a random effect with the type of stimuli as the random slope. Stimuli were entered as another random effect.

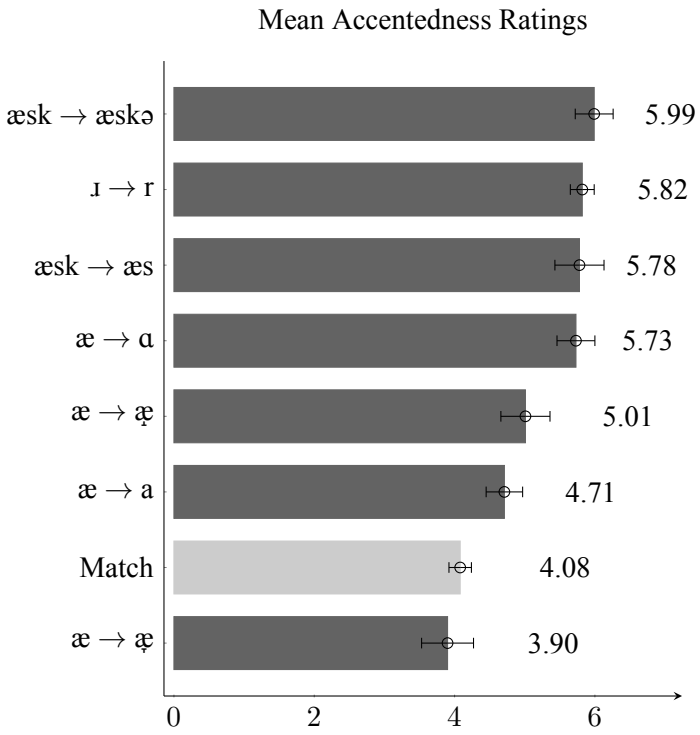

Figure 4.4: Mean Accentedness Ratings of Stimuli in "*Ask her*"

Model comparisons using likelihood ratio tests revealed that the type of stimuli contributed significantly to model fit ($\chi 2 = 18.40$, $p < .01$), showing that the eight types of stimuli were indeed rated differently. Trial number was another significant contributing factor to model fit ($\chi 2 = 7.84$, $p < .01$), showing that the same stimulus would be considered more accented if it occurred late in the experiment. The interaction between type of stimuli and trial numbers did not contribute significantly to model fit ($\chi 2 = 5.04$, $p = .66$), showing that rating differences of the stimuli were consistent overtime. To further investigate the rating differences between the eight type of stimuli, Helmert-contrast-coding was implemented to create seven contrasts (Table 4.2).

Helmert-contrast coding was used to achieve comparisons among the eight types of stimuli. For

Table 4.2: Stimuli Contrasts

Levels	Contrast1	Contrast2	Contrast3	Contrast4	Contrast5	Contrast6	Contrast7
æ→a̟	-0.5	-0.333	-0.25	-0.2	-0.167	-0.143	-0.125
Match	0.5	-0.333	-0.25	-0.2	-0.167	-0.143	-0.125
æ→a	0	0.667	-0.25	-0.2	-0.167	-0.143	-0.125
æ→a̟	0	0	0.75	-0.2	-0.167	-0.143	-0.125
æ→ɑ	0	0	0	0.8	-0.167	-0.143	-0.125
æsk→æs	0	0	0	0	0.833	-0.143	-0.125
ɹ→r	0	0	0	0	0	0.857	-0.125
æsk→æskə	0	0	0	0	0	0	0.875

example, the first contrast in Table 4.2 (i.e., Contrast1) compares ratings of [æ]→[a̟] to ratings of the match stimuli, while the second contrast (i.e., Contrast2) compares ratings of [æ]→[a] to ratings of both [æ]→[a̟] and the match stimuli. Since the seven contrasts were entered into the model as fixed effect factors, model comparisons using the likelihood ratio test could show whether any of the contrasts contributed significantly to model fit.

The full model took accentedness ratings as the dependent variable, the seven contrasts as fixed effects. Trial number and the interactions between trial number and the seven contrasts were also entered as fixed effects. Raters and stimuli were used as random effects. Model comparisons were achieved by excluding one contrast from the full model, and then comparing the new model to the full model, using the likelihood ratio test.

The results show the first contrast did not contribute significantly to model fit ($\chi2 = 0.02$, $p = .90$), indicating that the ratings of [æ]→[a̟] did not differ significantly from ratings of the match stimuli. The second contrast did not contribute significantly to model fit ($\chi2 = 0.78$, $p = .38$), indicating that the ratings of [æ]→[a] did not differ significantly from ratings of both [æ]→[a̟] and the match stimuli. With the current coding scheme, only the fifth contrasts contributed significantly to model fit ($\chi2 = 8.04$, $p < .01$), showing that ratings of [æsk]→[æs] were, in general, higher than ratings of vowel mismatches and the match stimuli.

To further achieve pairwise comparisons among the eight types of stimuli, reference levels for

Helmert-contrast coding were manipulated in several different ways. For example, in order to directly investigate whether ratings of [æ]→[ɑ] were higher than ratings of the match stimuli, the coding scheme was rearranged, as shown in Table 4.3. The first contrast now directly compares ratings of [æ]→[ɑ] to ratings of the match stimuli.

Table 4.3: Rearranged Stimuli Contrasts

Levels	Contrast1	Contrast2	Contrast3	Contrast4	Contrast5	Contrast6	Contrast7
æ→ɑ	-0.5	-0.333	-0.25	-0.2	-0.167	-0.143	-0.125
Match	0.5	-0.333	-0.25	-0.2	-0.167	-0.143	-0.125
æ→ɐ̞	0	0.667	-0.25	-0.2	-0.167	-0.143	-0.125
æ→a	0	0	0.75	-0.2	-0.167	-0.143	-0.125
æ→ɐ̟	0	0	0	0.8	-0.167	-0.143	-0.125
æsk→æs	0	0	0	0	0.833	-0.143	-0.125
ɹ→r	0	0	0	0	0	0.857	-0.125
æsk→æskə	0	0	0	0	0	0	0.875

The full model was rebuilt using the rearranged seven contrasts as fixed effects. Trial number and the interactions between trial number and the contrasts were also fixed effects, while raters and the stimuli were entered as random effects. The dependent variable was still the accentedness ratings. Model comparisons using likelihood ratio tests show that the first contrast contributed significantly to model fit ($\chi 2 = 4.06$, $p < .05$), indicating that ratings of [æ]→[ɑ] are significantly higher than ratings of the match stimuli.

Contrasts were reconstructed in several other ways to allow pairwise comparisons of all the eight types of stimuli. The results are listed below in Table 4.4 where the "≫" symbol indicates significant differences. The type of stimuli on the left side of "≫" are more accented than the type of stimuli on the right side of "≫." The type of stimuli on the same side of the "≫" did not differ significantly from one another.

As shown in Table 4.4, coda deletion [æsk]→[æs], vowel paragoge [æsk]→[æskə], /ɹ/-trilling [ɹ]→[r] and vowel backing [æ]→[ɑ] received significantly higher rating than [æ]→[a], the match stimuli, vowel raising [æ]→[ɐ̟] and vowel lowering [æ]→[ɐ̞]. The diacritic marks of [ɐ̟] and [ɐ̞]

Table 4.4: Accentedness Ratings for "*Ask her*"

æsk→æs; æsk→æskə; ɹ→r; æ→ɑ	≫	æ→a; Match; æ→æ̟; æ→ æ̟;

means that these two pronunciations are sub-phonemic variations of /æ/. The results in Table 4.4 shows that these two sub-phonemic variations of /æ/ do not differ in accentedness. These results show that syllable structure mismatches and consonant mismatches were judged as being more accented than most vowel mismatches. Among the stimuli with vowel mismatches, only vowel backing [æ]→[ɑ] was judged as being more accented than the match stimuli. Ratings of other types of vowel mismatches are not significantly different from ratings of the match stimuli.

Context 2: "*Please call*"

Figure 4.5 summarizes the results for stimuli in the context of "*please call*." VOT-shortening on /p/ and /k/, vowel anaptyxis [pʰl]→[pʰəl] and coda devoicing [z]→[s] received realatively higher accentedness ratings than other types of mismatches. By comparison, vowel raising [ɑ]→[ɔ] and [ɑ]→[o], /l/-deletion in "*call*" and "*please*" and the match stimuli received relatively lower ratings.

Linear mixed-effects regression models were built to achieve pairwise comparisons between the different types of stimuli. Table 4.5 summarizes the results. There are three rankings for the different type of stimuli in the context of "*please call*." The ranking in the top row of Table 4.5 shows that VOT-shortening [pʰl]→[pl] and vowel anaptyxis [pʰl]→[pʰəl] were judged as being significantly more accented than vowel raising [ɑ]→[o], which was judged as being significantly more accented than the match stimuli, /l/-deletion in cluster [pʰl] and vowel raising [ɑ]→[ɔ].

A few types of stimuli are missing from the top row, because these stimuli did not receive either a significantly lower rating than [ɑ]→[o] or a significantly higher rating than the match stimuli. These stimuli are listed in the second and the third rows to specify their accentedness ratings in relation to other types of stimuli.

The second ranking in Table 4.5 shows that coda devoicing [z]→[s] and VOT-shortening [kʰ]→[k]

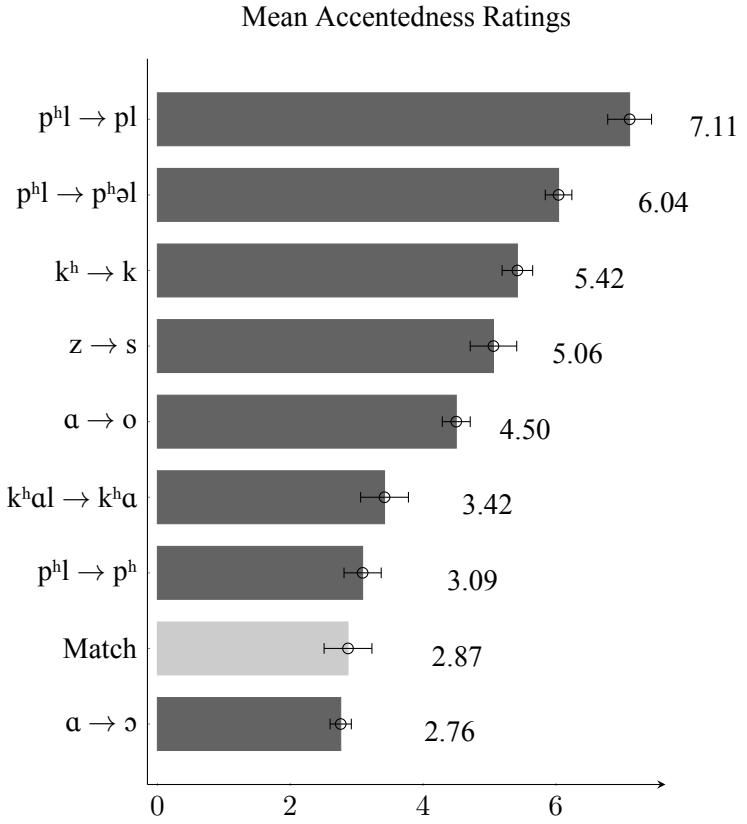

Figure 4.5: Mean Accentedness Ratings of Stimuli in "*Please Call*"

Table 4.5: Accentedness Ratings for "*Please call*"

pʰl→pl; pʰl→pʰəl	≫	ɑ→o	≫	Match; pʰl→pʰ; ɑ→ɔ
pʰl→pl; pʰl→pʰəl; z→s; kʰ→k	≫			kʰɑl→kʰɑ; Match; pʰl→pʰ; ɑ→ɔ
ɑ→o; z→s; kʰ→k (No significant difference)				

were not judged as significantly more accented than VOT-shortening [pʰl]→[pl] and vowel anaptyxis [pʰl]→[pʰəl], but were significantly more accented than /l/-deletion in "*call*" (i.e., [kʰɑl]→[kʰɑ]). The third ranking in Table 4.5 shows that ratings of vowel raising [ɑ]→[o],VOT-shortening [kʰ]→[k] and [z]→[s] did not differ significantly from one another.

Context 3: "*Six spoons*"

Figure 4.6 demonstrates the mean ratings of stimuli in the context of "*six spoons*." Consonant mismatches such as pronouncing coda /z/ as [ʃ] (i.e., [spũnz]→[spũnʃ]) and VOT-lengthening in /sp/ (i.e., [spũnz]→[sp^{h}ũn]) were judged to be more accented than other types of stimuli. Vowel mismatches, in general, were judged as relatively less accented.

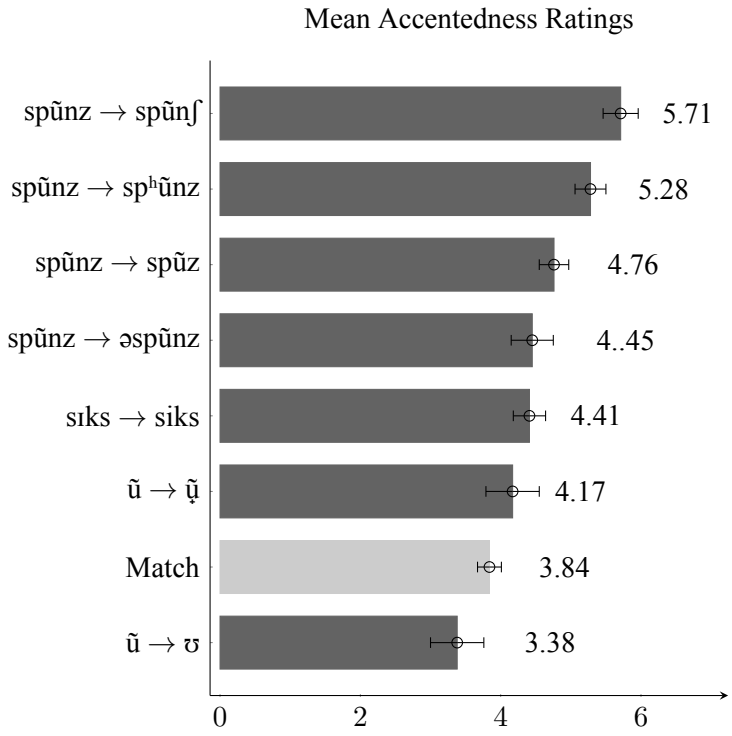

Figure 4.6: Mean Accentedness Ratings of Stimuli in "*Six spoons*"

Table 4.6 lists the three accentedness rankings for the different types of stimuli. Ratings of [spũnz]→[spũnʃ] were significantly higher than other types of mismatches. Ratings of the various types of vowel mismatches were not significantly different from the match stimuli. For syllable mismatches, /n/-deletion in "spoons" (i.e.,[spũnz]→[spũz]) was judged as being more accented than the match stimuli; prothesis in /sp/ (i.e., [spũnz]→[əspũnz]), however, did not receive a significant higher rating than the match stimuli.

Table 4.6: Accentedness Ratings for "*Six spoons*"

spũnz→spũnʃ	≫	spũnz→spʰũnz; spũnz→spũz	≫	ũ→ʊ; Match
spũnz→spũnʃ	≫	spũnz→spũz; spũnz→əspũnz; sɪks→siks; ũ→ṵ		
spũnz→spʰũnz	≫	spũnz→əspũnz; sɪks→siks; ũ→ṵ; ũ→ʊ; Match		

Context 4: "*Five thick*"

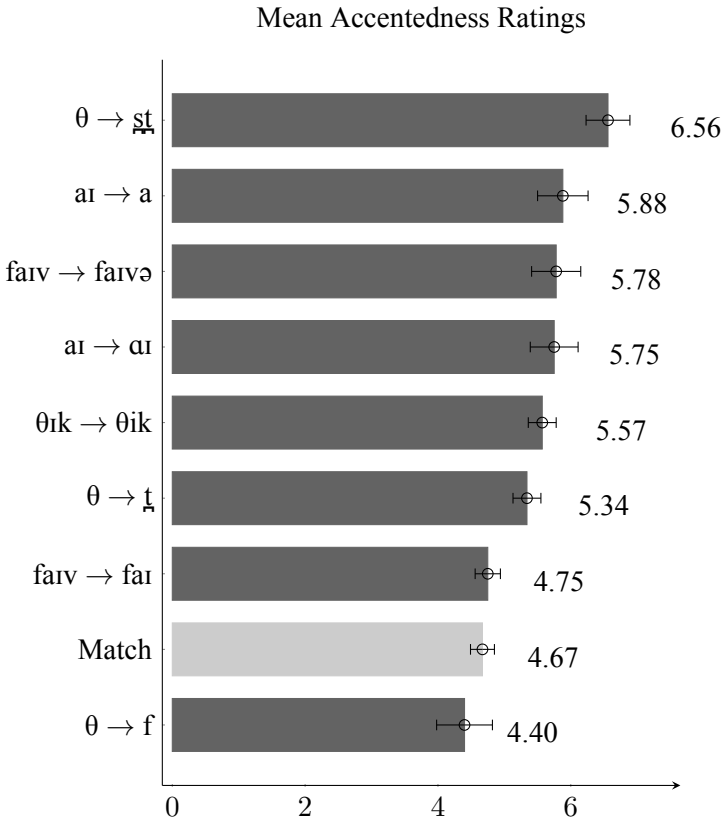

Figure 4.7: Mean Accentedness Ratings of Stimuli in "*Five thick*"

Figure 4.7 demonstrates the mean ratings of stimuli in the context of "*five thick*." Consonant mismatch [θ]→[st] received the highest ratings. This type of mismatch was classified by the SAA as a type of consonant variation. The current study followed the SAA classification. However, changing

/θ/ to /st/ also alters syllable structure. In other words, the L2 production [s̪t̪ɪk] contains both a consonant mismatch and a syllable structure mismatch. Therefore, it was expected that [θ]→[s̪t̪] would receive a higher mean accentedness rating.

Table 4.7 shows that [θ]→[s̪t̪] received significantly higher accentedness ratings than other types of stimuli. Off-glide deletion [aɪ]→[a], vowel paragoge [faɪv]→[faɪvə] and vowel tensing [θɪk]→[θik] were judged as significantly more accented than [θ]→[t̪], the match stimuli, and [θ]→[f].

Table 4.7: Accentedness Ratings for "*Five thick*"

θ→s̪t̪	≫	aɪ→a; faɪv→faɪvə ; θɪk→θik	≫	θ→t̪; faɪv→faɪ; Match; θ→f

As shown in Table 4.8, the most accented stimuli were the ones with a non-English phoneme (i.e., [smɑl→smɑ[], [pʰl→pʰ[]). Alternations between English phonemes were not as accented.

Context 5: "*Small plastic*"

Figure 4.8 demonstrates the mean ratings of stimuli in the context of "*small plastic*." Consonant mismatches [pʰl]→[pʰɾ],[smɑl]→[smɑ[] and [pʰl]→[pʰ[] were rated as the most accented. Vowel mismatches were relatively less accented. /t/-deletion in "*plastic*" (i.e., [pʰlæstɪk]→[pʰlæsɪk]) received almost the same mean rating as the match stimuli.

Table 4.8 shows that ratings of [pʰl]→[pʰɾ], [smɑl]→[smɑ[] and [pʰl]→[pʰ[] were significantly higher than other types of stimuli. Ratings of VOT-shortening in "*plastic*" (i.e., [pʰl]→[pl]), vowel tensing in "*plastic*" (i.e.,[ɪ]→[i]), consonant voicing in "*small*" (i.e., [sm]→[zm]) and various vowel mismatches are not significantly higher than ratings of the match stimuli.

4.6.5 Effects of Phonological Context

The analysis above shows that stimuli with consonant mismatches were rated as generally more accented than stimuli with vowel mismatches in all five contexts. However, phonological contexts also seem to have affected raters' accentedness judgment.

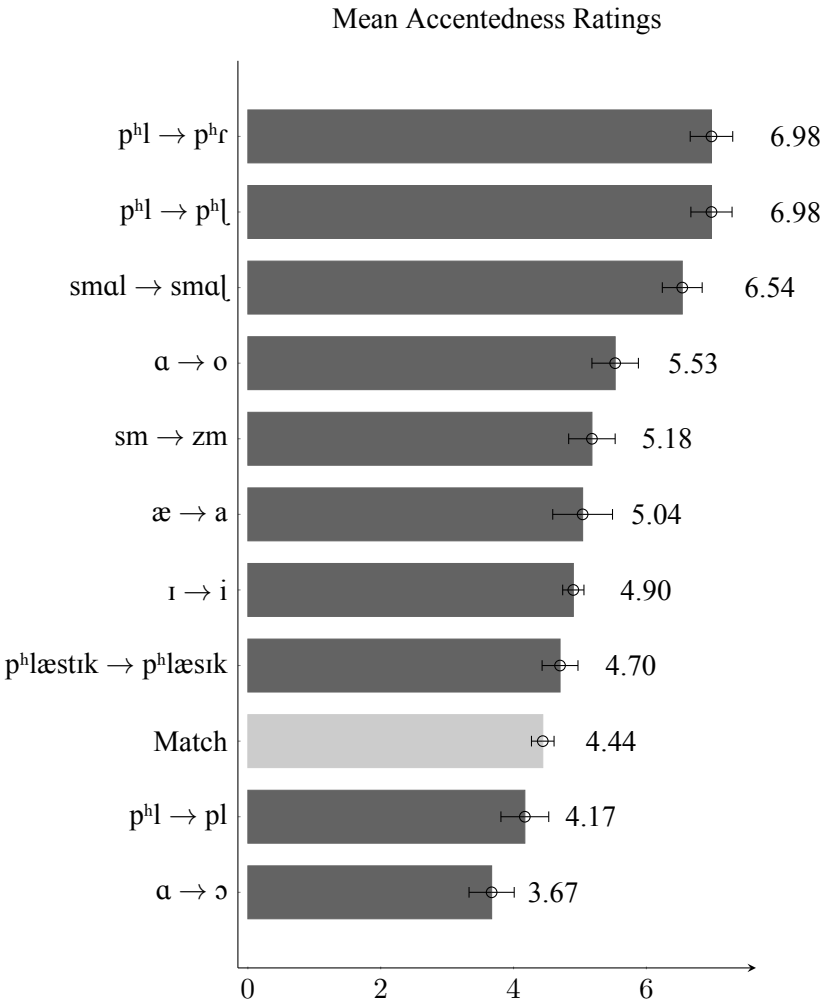

Figure 4.8: Mean Accentedness Ratings of Stimuli in "*Small plastic*"

Table 4.8: Accentedness Ratings for "*Small plastic*"

smɑl→smɑɭ; pʰl→pʰɭ; pʰl→pʰɾ	≫	pʰl→pl; ɑ→ɔ; æ→a; Match; ɪ→i; sm→zm; pʰlæstɪk→pʰlæsɪk; ɑ→o;

For example, the accentedness of VOT-shortening might be affected by where the shortening happened, as illustrated in Figure 4.9, where the * marks the statistically significant difference between accentedness ratings of a given stimulus (e.g., [pl]) and its L1 target form (e.g., [pʰl]). VOT shortening in "*please call*" (i.e., [pʰl]→[pl]) was assigned higher ratings than the match stimuli. However, ratings of VOT shortening in "*small plastic*" (i.e., [pʰl]→[pl]) and in "*call*" (i.e., [kʰ]→[k]) was not significantly higher than the match stimuli.

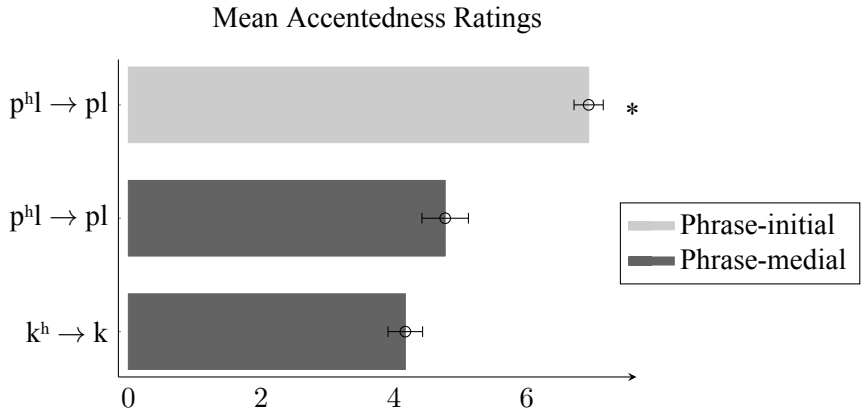

Figure 4.9: Mean Accentedness Ratings of VOT shortening

The effect of phonological context was also observed on the accentedness of vowel mismatches. Figure 4.10 shows the accentedness ratings of vowel tensing (i.e., [ɪ]→[i]) in three contexts. Only vowel tensing in "*thick*" was rated as more accented than the match stimuli.

Similar effects of phonological context have been found on syllable mismatches (Figure 4.11. Coda deletion is often allowed in L1 speech. In most contexts, coda deletion was indeed rated to be less accented than other types of mismatches (e.g., [faɪv]→[faɪ] in context "*five thick*"). Interestingly, obstruent coda deletion in "*ask her*" (i.e., [æsk]→[æs]) was rated as accented, showing that raters were sensitive to the phonological context where coda deletion could happen.

Stimuli with anaptyxis (i.e., [pʰl]→[pʰəl]) and stimuli with paragoge (i.e., [æsk]→[æskə]) were rated as more accented than stimuli with consonant mismatches in their respective contexts. Prothesis of /sp/ was not rated as accented as anaptyxis or paragoge (See Figure 4.12). These results show

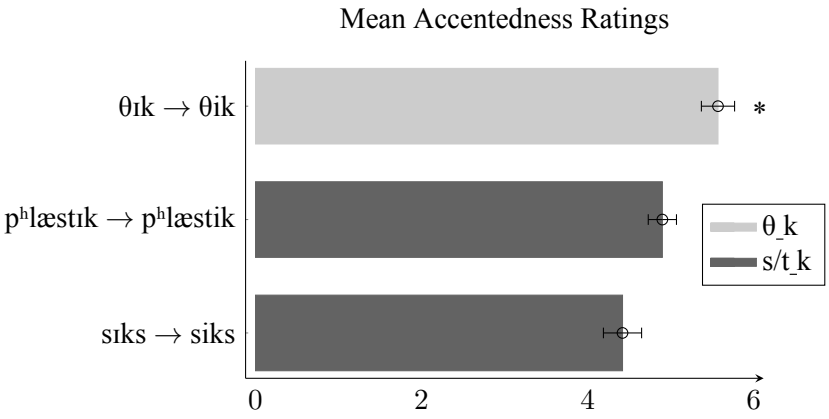

Figure 4.10: Mean Accentedness Ratings of Vowel Tensing

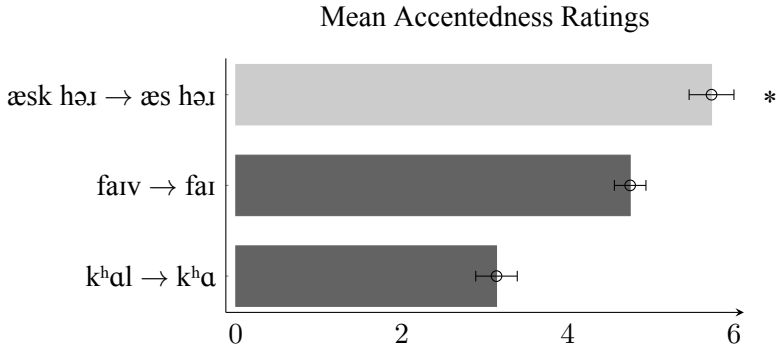

Figure 4.11: Mean Accentedness ratings of Coda Deletion

that the effect of syllable mismatches on accentedness concerns both the specific type of mismatches and the phonological contexts the mismatches are in.

4.6.6 Summary

The first part of the analyses focused on ratings of four types of stimuli, namely stimuli with consonant mismatches, vowel mismatches, syllable structure mismatches, and stimuli without mismatches (i.e., the match stimuli). The results show that stimuli with consonant mismatches were rated as more

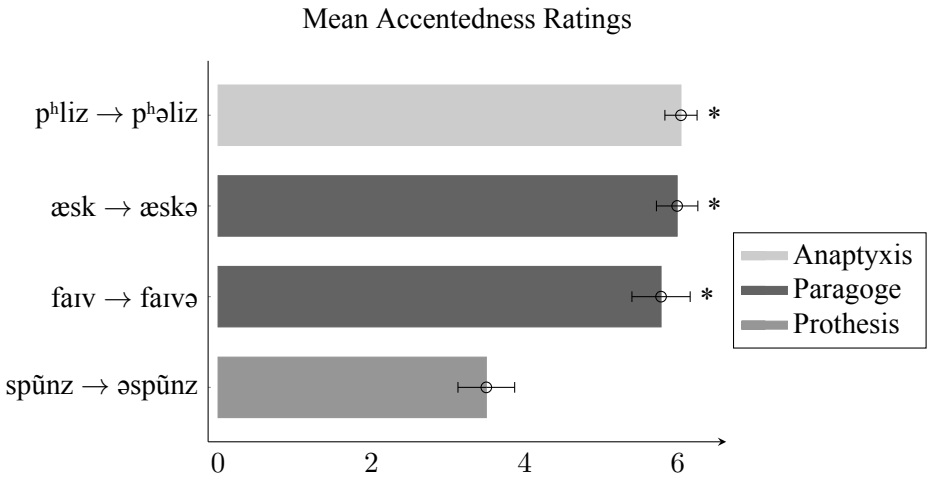

Figure 4.12: Mean Accentedness Ratings of Vowel Epenthesis

accented than stimuli with vowel and syllable structure mismatches, which in turn were rated as more accented than the match stimuli. During the entirety of the experiment, consonant mismatches were always rated higher than other types of mismatches. The match stimuli always received lower accentedness ratings. Syllable structure and vowel mismatches were always rated lower than consonant mismatches and higher than the match stimuli. Further analysis show that accentedness ratings of the same type of mismatches may vary depending on the phonological context of the mismatches.

4.7 Discussion

As shown in the analysis on individual mismatches, the most accented stimuli are the ones with a non-English sound (e.g., retroflex [ɭ], and trill [r]). By comparison, the alternations between English consonant phonemes were rated as relatively less accented (i.e., [θ]→[f]). The judgments on the alternation betwen sounds of the same phoneme were not as clear. For example, the effect of VOT-shortening (e.g., [pʰl]→[pl]) on accentedness judgments is more prominent phrase-initially than phrase-medially. The reason for such a phenomenon could be attributed to L1 listeners' sensitivity to the existence/absence of the domain-initial strengthening effect on domain-initial

aspirated plosives, which might also account for conflicting findings on the accentedness of VOT-shortening/lengthening in the previous literature (González-Bueno, 1997; Magen, 1998; Riney and Takagi, 1999).

The effect of vowel mismatches on accentedness judgment is not as clear as that of consonant mismatches. Several reasons might account for the mixed findings presented here. First, accentedness of some vowel mismatches was also affected by phonological contexts. Second, vowel quality change might often be perceived as dialectal rather than foreign accented. Depending on the raters' own dialects and their exposure to other varieties of English, many types of mismatches could be native-like. The current study classified it as a case of vowel mismatch simply because it does not match the most common L1 production [kʰɑl]. The relative lower accentedness ratings of [ɑ]→[ɔ] indeed show that the raters probably judged [kʰɔl] as a relatively more native-like production. Since most of the consonant mismatches are not L1 dialectal variations (except [θ]→[t] and [θ]→[t]), the general claim that consonant mismatches are more accented than vowel mismatches could be misleading. To further investigate this issue, the types of mismatch should be examined so that dialectal variations do not skew the results. Chapter 6 discusses an experiment that resolves this issue.

Although stimuli with syllable mismatches were judged as less accented than stimuli with consonant mismatches, and more accented than the match stimuli, different types of syllable mismatches seem to affect accentedness rating differently. For example, stimuli with anaptyxis (i.e., [pʰl]→[pʰəl]) and paragoge (i.e., [æsk]→[æskə]) were rated as being more accented than the match stimuli in their respective contexts. Prothesis of /sp/ was not as accented as the other two types of epenthesis. The reason could be attributed to the similar sonority profile between the prothesized *s*-cluster (i.e., [əsp]) and the L1 target (i.e., [sp]). As Gouskova (2001) claims, prothesis of *s*-clusters exhibits a falling sonority profile, which does not alter the sonority profile of the original *s*-cluster. Anaptyxis and paragoge of consonant clusters, on the other hand, change the sonority profile of the original clusters. Raters' sensitivity to sonority profile could have affected their accentedness judgment.

Consonant deletions also exhibit a different degree of impact on accentedness judgment. Coda

/v/ deletion in "*five thick*" and coda /l/ deletion in "*please call*" did not contribute much to accentedness ratings. However, coda /k/ deletion in "*ask her*" was judged as being relatively more accented. These results show that the accentedness of syllable mismatches associates with both the specific type of mismatches and the phonological contexts the mismatches are in.

4.8 Limitations

Experiment 1 has several limitations. First, no training session was presented to the raters. Since the raters were not aware of the full range of accentedness at the beginning, most of the ratings were around five (i.e., the middle point on a 9-point scale), showing that the raters were not committing themselves to extreme opinions. As the experiment progressed, more native-sounding stimuli (i.e., the match stimuli) were being heard, which could have affected the accentedness perception of other types of L2 stimuli. Such a result is consistent with findings in Flege and Fletcher (1992), which showed that the proportion of native (or near-native) stimuli in an experiment positively correlates with the perceived accentedness of L2 stimuli. It is therefore necessary to include a training session that could familiarize raters with the full range of accents presented in the experiment.

Second, raters of Experiment 1 could only listen to the stimuli, without knowing what the speakers were trying to say. In other words, intelligibility could have affected accentedness ratings. For example, pronouncing "*six*" as [siks] could indeed be considered native-like, since "*seeks*"([siks]) is an English word. Raters could only realize the intended meaning of each stimulus as the experiment progressed, which might have contributed to the increase of ratings during the whole experiment. It is therefore necessary for future experimental design to separate intelligibility from accentedness.

Third, the analysis presented in this chapter did not examine sub-phonemic acoustic information, which could have biased the results. For example, VOT-shortening in phrase-initial environments were rated as more strongly accented than VOT-shortening in phrase-medial environments. The current analysis attributes this phenomenon to contextual effect without concerning how much the VOTs were shortened. Alternative arguments could indeed speculate on whether the degree of VOT-shortening could correlate with the degree of accentedness. As shown in Goldinger (1998) and Nielsen (2011), acoustic signals that were thought as idiosyncratic (e.g., sub-phonemic VOT

duration) could potentially be perceivable, raising questions on whether and how sub-phonemic information affects accentedness judgments.

Experiment 1 also found that prothesis of /sp/ was rated as less accented than the other two types of vowel epenthesis. The reason was attributed to the sonority profile of the prothesized *s*-cluster. Another reason for why prothesis of *s*-clusters was rated as less accented could potentially lie in the saliency of the epenthetic vowel. That is, the prothesized vowel could be very short in duration, making the vowel less noticeable to raters.

Experiment 1 classified phonetic patterns in L2 speech based on whether they match their L1 target productions. The most common L1 productions were considered as the L1 target productions. Such treatment was based on the assumption that L1 listeners would be familiar with the most common L1 productions and consequently consider the most common L1 productions as exhibiting "no foreign accent at all" (i.e., a rating of 1 on the 9-point scale).

Some of the mismatches, although they did not match the most common L1 productions, did match some L1 productions. For example, among the 100 L1 speakers surveyed by the current study, one person pronounced "*ask*" as [æs]. As discussed previously, the word "*ask*" occurred in the context of "*ask her*." [k]-dropping is not usually allowed in this context. Since there is at least one L1 speaker who dropped the [k], we did not term [æs] as an "L2" variation of "*ask*" in the context of "*ask her*." The question for the current study is instead whether uncommon and non-dialectal productions such as the [æs] would be judged as foreign accented. Results from Experiment 1 seem to have provided an affirmative answer. However, Experiment 1 did not carefully examine the relationship between accentedness and the frequency of occurrence of a phonetic pattern in L1 speech. A more detailed study is needed to further explore this issue.

This dissertation addresses these limitations in Experiment 2 (Chapter 5) and Experiment 3 (Chapter 6).

Chapter 5: Experiment 2

5.1 Introduction

Experiment 1 (Chapter 4) of the current study investigates the accentedness of various speech patterns in non-native (L2) English speech. The most common native (L1) productions were defined as the L1 target productions. L2 speech that differs from its L1 target production was considered as having mismatches. Results of Experiment 1 show that L2 stimuli with consonant mismatches were judged as being more accented than L2 stimuli with syllable and vowel mismatches. Raters of Experiment 1 did not receive any training before the experiment and were not aware of the intended meaning of each stimulus. These experimental limitations could have affected raters' accentedness judgments. In addition, sub-phonemic acoustic information of the stimuli was not examined. Questions remain as to whether and how the experimental limitations and sub-phonemic information of the stimuli affected accentedness perception. Experiment 2 of the current study aims to address the potential shortcomings of Experiment 1.

The same 100 stimuli from Experiment 1 were used in Experiment 2. An additional 150 raters were recruited on Amazon Mechanical Turk (MTurk). In addition to ranking the accentedness of different types of mismatches, Experiment 2 also investigates the effect of sub-phonemic acoustic information on accentedness perception.

5.2 Procedure

To familiarize raters with the full range of accents, a training phase was added at the beginning of the experiment, consisting of ten trials. Ten stimuli were selected for the training phase based on the accentedness ratings collected in Experiment 1. Raters of Experiment 1 judged the accentedness of a stimulus on a 9-point Likert-like scale. Among the ten stimuli, two received the highest accentedness

ratings in Experiment 1, two received the lowest ratings, two received an average rating at around 6, two received an average rating at around 5, another two received an average rating at around 4. Table 5.1 illustrates the ten stimuli and their mean accentedness ratings obtained in Experiment 1. To familiarize raters with the five contexts, each context occurred twice during the training phase. The same ten stimuli were used during the training phase for all raters. The presentation of the ten training phase stimuli was randomized for each rater.

Table 5.1: Mean Ratings of the Ten Training Session Stimuli

Contexts	Mean Ratings	Types of Stimuli
please call	2.22	Match
ask her	3.61	Match
small plastic	3.96	Vowel Mismatch
six spoons	4.10	Syllable Mismatch
five thick	4.89	Vowel Mismatch
five thick	5.07	Syllable Mismatch
small plastic	5.73	Syllable Mismatch
six spoons	5.71	Consonant Mismatch
please call	6.10	Syllable Mismatch
ask her	7.27	Consonant Mismatch

To control for the effect of intelligibility, the script of each stimulus was shown on the screen together with the instructions. The raters were instructed to click on a button to hear each stimulus, after which a 9-point Likert-like scale would appear. Raters then rated the accentedness of the stimulus and moved on to the next trial.

After the training phase, raters were informed that the actual experiment would begin. This section of the experiment (the testing phase, henceforth) consisted of 100 trials. The presentation of the 100 trials was randomized in the same manner as in Experiment 1. The stimuli used for the testing phase were the same 100 speech samples used for Experiment 1, which included the ten stimuli used for the training phase. After the 100 trials, raters filled out the same demographic questionnaire as in Experiment 1. The maximum time allowed to complete the experiment was 40 minutes. The raters could only listen to each stimulus once. Figure 5.1 demonstrates the interface of the experiment.

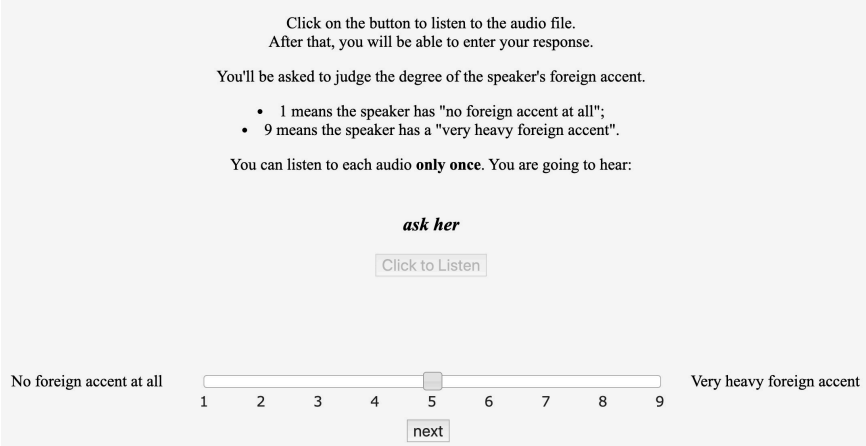

Figure 5.1: Interface of Experiment 2

5.3 Raters

150 raters were recruited on the MTurk platform. Among the 150 raters, 17 raters self-reported as being an L2 English speaker, or having speech or hearing related problems, or being born outside of the United States. These 17 raters' responses were excluded from the final analysis. The remaining 133 raters self-reported as L1 English speakers who were born and currently reside in the United States. Among the 133 raters, 68 were male, 58 were female; seven raters did not report their gender[1]. The mean age of the 133 raters was 38.42 (SD = 11.84). The raters were from 33 states and the District of Columbia in the continental United States (See Appendix B for rater demographics). The average time for the 133 raters to complete the experiment was 15.96 minutes (SD = 5.47 minutes). All 150 raters were paid $0.50 for their participation.

[1]Some of these raters conveyed to the experimenter their dissatisfaction with the male/female dichotomy implied by the questionnaire.

5.4 Results

5.4.1 Experiment 2: Predictions

Results of Experiment 1 show that ratings of the 100 stimuli were not consistent over time. Detailed analysis show that raters were making adjustments during the first few trials of the experiment, probably due to their unfamiliarity with the procedure of the experiment and the range of accents. Since Experiment 1 did not inform raters of the intended meaning of each stimulus, intelligibility of the stimuli could have affected raters' judgment. Experiment 2 accounted for these issues by including a training phase and directly informing raters of the intended meaning of each stimulus. It is, thus, expected that the ratings during the testing phase of Experiment 2 would be more consistent than ratings of Experiment 1.

Experiment 1 shows that stimuli with consonant mismatches were judged to be more accented than stimuli with vowel and syllable mismatches. Experiment 1 also shows that phonological context potentially affected how accented a stimulus was judged. The same results are expected for Experiment 2.

5.4.2 Segmental and Structural Mismatches

Ratings from the training phase were excluded from the analysis. Only ratings during the testing phase were analyzed. All the stimuli were rated on a scale from 1 to 9: the larger the number, the more foreign-accented a stimulus was judged. As mentioned in previous chapters, the most common L1 productions in the SAA were considered L1 target productions. 25 of the 100 L2 speech samples were transcribed the same as their respective L1 target productions. These 25 speech samples were termed the match stimuli. IPA transcriptions of the remaining 75 stimuli were not the same as their respective L1 target productions. These 75 stimuli were termed the mismatch stimuli. The mismatch stimuli were further grouped into four types to reflect how they differ from L1 target productions.

There were 5 phonological contexts and 4 types (i.e., consonant vs. vowel vs. syllable vs. match) for the 100 stimuli. There were 20 stimuli for each context, five of which contained only one consonant mismatch, five contained only one vowel mismatch, five contained only one syllable

structural mismatch, and another five did not contain any segmental or syllable mismatches (i.e., the match stimuli).

The mean ratings of all of the 100 stimuli was 5.13 (SD=2.29). Raters assigned higher ratings to stimuli with segmental and syllable structural mismatches (M=5.36, SD=2.21) than the match stimuli (M=4.42, SD=2.35). Ratings of stimuli with consonant mismatches (M=5.70, SD=2.11) were on average higher than ratings of stimuli with syllable mismatches (M=5.40, SD=2.29), which was on average higher than stimuli with vowel mismatches (M=5.00, SD=2.18). The general trend is similar to results of Experiment 1. Figure 5.2 demonstrates the mean ratings of each type, where the error bars represent the 95% confidence intervals.

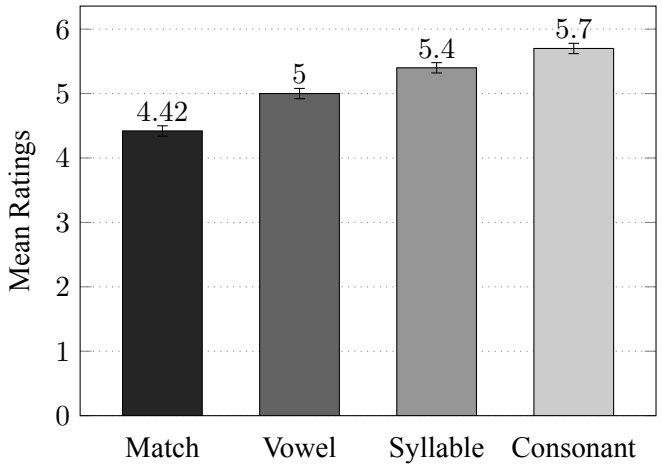

Figure 5.2: Mean Ratings by Type of Stimuli on the Scale from 1 to 9

Linear mixed-effects regression models were employed with the *lme4* package in R (Bates et al., 2014) to investigate segmental and syllable influences on foreign accent perception. The regression models were built the same way as the ones used for Experiment 1. Model comparisons were conducted using the Likelihood Ratio Test as described in Chapter 4. The results show that the DTW scores, which represent prosodic differences between L1 and L2 speech samples, and the interactions involving the DTW scores did not contribute significantly to model fit, showing that prosodic information of the stimuli might not be a major contributing factor to accentedness judgment.

Just as in Experiment 1, types of stimuli were coded using Helmert contrasts. Results show that the contrast between consonant and syllable mismatches did not achieve significant contribution to model fit ($\chi2$=1.60, p=.21), indicating that ratings of stimuli with consonant and syllable mismatches did not differ significantly from each other. The second contrast, which compared vowel mismatches with consonant and syllable mismatches, contributed significantly to model fit ($\chi2 = 5.15$, $p < .02$), showing that stimuli with consonant and syllable mismatches were rated as being more accented than stimuli with vowel mismatches. The third contrast, which compares the match stimuli with the three types of mismatch stimuli contributed significantly to model fit ($\chi2 = 19.84$, $p < .001$), showing that stimuli with mismatches were rated as being more foreign-accented than the match stimuli.

These results suggest that all three types of mismatches contributed to perceived foreign-accentedness. Among the three types of mismatches, vowel mismatches were rated to be the least accented.

5.4.3 Ratings across Time

Just as in Experiment 1, trial number, which represented time, contributed significantly to model fit (β=0.6, $\chi2 = 69.91$, $p < .001$), while the interactions between trial number and the three contrasts did not contribute significantly to model fit. These results show that ratings of the four types of stimuli increased over time.

Unlike Experiment 1, Experiment 2 included a training phase, containing ten stimuli covering the range of accents included in the experiment. Ratings obtained by Experiment 1 show that the raters were making adjustments during the first few trials. To investigate whether the inclusion of a training phase and the controlling of intelligibility had affected accentedness judgment, a Smoothing Spline ANOVA (SSANOVA) method was implemented with the *gss* package in R (Gu, 2014). The results are shown in Figure 5.3, where the solid line represents ratings from Experiment 1 and the dotted line represents ratings from Experiment 2. The shaded areas represent 95% Bayesian confidence intervals. As Figure 5.3 shows, ratings of Experiment 1 experienced a sudden drop during the first 10 to 15 trials and then the ratings went up. However, ratings obtained by Experiment 2 were more consistent. In other words, no adjustment was observed during the testing phase of Experiment 2. Ratings of Experiment 2 were also generally higher than ratings of Experiment 1.

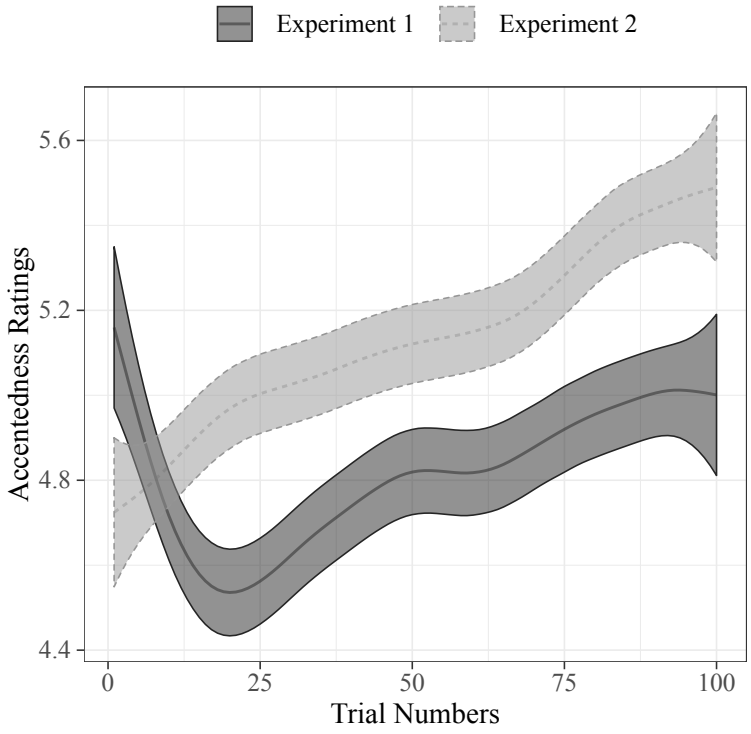

Figure 5.3: Ratings across Time (Experiment 1 and Experiment 2)

The ten training phase stimuli were also used in the testing phase. It is possible that the raters' judgments on these ten stimuli during the testing phase are different from their judgments on other stimuli, simply because the ten stimuli were heard during the training phase. Linear mixed-effects model was run to investigate whether ratings on the ten stimuli during the testing phase are different from ratings on the remaining 90 stimuli. The ratings were the dependent variable.The times a stimulus was heard (once vs. twice), the four types of stimuli, trial number and the interactions between these factors were included as fixed effects. The raters and the stimuli were entered as two random effects. The results show that the times a stimulus was heard did not contribute significantly to model fit ($\chi 2 = 0.01$, $p = .93$). Therefore, ratings of the ten stimuli during the testing phase might not have been affected by the fact that the ten stimuli were included the training phase.

5.4.4 Individual Mismatches

The analysis above focuses on broad categories such as consonant, vowel, and syllable mismatches. However, there were 11 types of consonant mismatches, five types of vowel mismatches and two types of syllable mismatches involved in the experiment. Although consonant mismatches appeared in general to be more accented than the other two types of mismatches, it might be too hasty to draw the conclusion that all consonant mismatches are more accented than the other two types. This section presents a more detailed analysis, which focuses on individual mismatches within a given phonological context. The following section first discusses the statistical methods using one phonological context as an example. Since statistical methods used for the other four phonological contexts are the same as methods used for the first context, the details for these four contexts are omitted.

Context 1: "*Ask her*"

For the context "*ask her*," syllable mismatches such as consonant deletion (i.e., [æsk]→[æs]) and vowel epenthesis (i.e., [æsk]→[æskə]) received the highest ratings (i.e., most accented), while vowel mismatches received relatively lower ratings. Figure 5.4 demonstrates the mean ratings of the seven individual mismatches and the match stimuli, where the error bars represent 95% confidence intervals.

Ratings in Figure 5.4 showed that consonant deletion (i.e., [æsk]→[as]) and vowel epenthesis (i.e., [æsk]→[æskə]) received the highest ratings, while the match stimuli and vowel lowering (i.e., [æ]→[ᵆ]) received the lowest ratings. Vowel backing (i.e., [æ]→[ɑ]) and vowel raisings (i.e., [æ]→[ᵆ]) received relatively higher ratings than vowel lowering. The mean rating for consonant change (i.e., [ɹ]→[r]) was relatively lower than the two syllable mismatches, but higher than all the vowel mismatches. To evaluate whether the rating differences observed in Figure 5.4 are statistically significant, mixed-effects linear regression models were constructed to investigate the effect of individual mismatches on accentedness.

Ratings were the dependent variable. The eight types of stimuli and trial number were entered as fixed effects. Raters were entered as a random effect with types of stimuli as the random slope.

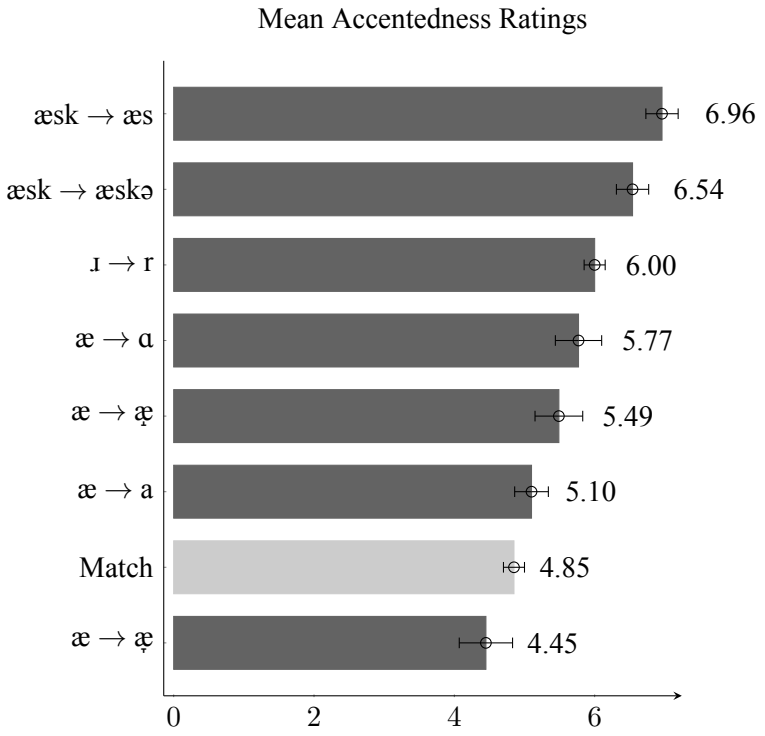

Figure 5.4: Mean Accentedness Ratings of Stimuli in "*Ask her*"

Stimuli were entered as another random effect. DTW scores were not included in the model since analysis in the previous section found no evidence of the effect of DTW on accentedness ratings. Model comparisons using likelihood ratio tests revealed that type of stimuli contributed significantly to model fit ($\chi 2 = 26.42$, $p < .001$), showing that the eight types of stimuli were indeed rated differently. Trial number was another significant contributing factor to model fit ($\beta=0.03$, $\chi 2 = 10.20$, $p < .001$), showing that the same stimulus would be considered more accented if it occurred late in the experiment. The interaction between the type of stimuli and trial numbers did not contribute significantly to model fit ($\chi 2 = 6.28$, $p =.51$), showing that rating differences between the four types of stimuli were consistent overt ime.

To further investigate pairwise rating differences between the eight type of stimuli, Helmert-contrast-coding was implemented to create seven contrasts (Table 5.2), which were entered into another mixed-effects model as seven fixed effects. Trial number and the interactions between trial

number and the seven contrasts were also entered as fixed effects. Raters and stimuli were used as random effects.

Table 5.2: Stimuli Contrasts

Levels	Contrast1	Contrast2	Contrast3	Contrast4	Contrast5	Contrast6	Contrast7
æ→ᵆ	-0.5	-0.333	-0.25	-0.2	-0.167	-0.143	-0.125
Match	0.5	-0.333	-0.25	-0.2	-0.167	-0.143	-0.125
æ→a	0	0.667	-0.25	-0.2	-0.167	-0.143	-0.125
æ→ᵆ	0	0	0.75	-0.2	-0.167	-0.143	-0.125
æ→ɑ	0	0	0	0.8	-0.167	-0.143	-0.125
æsk→æs	0	0	0	0	0.833	-0.143	-0.125
ɹ→r	0	0	0	0	0	0.857	-0.125
æsk→æskə	0	0	0	0	0	0	0.875

Model comparisons were achieved using likelihood ratio tests as described previously in Chapter 4. The results show that the fifth, the sixth and the seventh contrasts contributed significantly to model fit, while the other four contrasts did not contribute significantly to model fit. These results show that syllable mismatches (i.e., [æsk]→[æskə], [æsk]→[æs]) and /ɹ/-trilling in "her" were more accented than other types of stimuli. Contrasts were reconstructed in several different ways to allow pairwise comparisons of different types of stimuli.

The results are listed in Table 5.3 where the "≫" symbol indicates significant differences. The types of stimuli on the left side of "≫" were judged as being more accented than types of stimuli on the right side of "≫". The types of stimuli on the same side of the "≫" did not differ significantly from one another.

Table 5.3: Accentedness Ratings for "*Ask her*"

æsk→æs	≫	æsk→æskə	≫	æ→a; Match; æ→ᵆ;
æsk→æs	≫	æ→ᵆ	≫	Match; æ→ᵆ
æ→ᵆ; æ→ɑ; ɹ→r; æ→a (no significant difference)				

Table 5.3 lists three rankings. The first row shows that ratings of [æsk]→[æs] were significantly higher than ratings of [æsk]→[æskə], which were significantly higher than ratings of [æ]→[a], the match stimuli or [æ]-[ᵆ].

[æ]→[ɑ] and [ɹ]→[r] are missing from the first row, because ratings of these two stimuli were not significantly lower than ratings of [æsk]→[æskə] or significantly higher than ratings of [æ]→[a]. Therefore, [æ]→[ɑ] and [ɹ]→[r] could not be placed at any side of the "≫"s. [æ]→[ᵆ] was also missing from the first row, because its ratings were not significantly higher than [æ]→[a], but significantly higher than ratings of the match stimuli or [æ]→[ᵆ].

Two more rankings were therefore created to specify accentedness rankings regarding [æ]→[ɑ], [ɹ]→[r] and [æ]→[ᵆ]. The second row lists the second ranking that shows that ratings of these three stimuli were significantly lower than ratings of [æsk]→[æs], but significantly higher than ratings of the match stimuli or [æ]→[ᵆ]. The third row lists the third ranking, showing that ratings of [æ]→[ᵆ] were significantly lower than [æsk]→[æskə]. The fourth row shows that ratings of [æ]→[ᵆ], [æ]→[ɑ], [ɹ]→[r], [æ]→[a] did not differ significantly from one another.

The rankings above show that [æsk]→[æs] and [æsk]→[æskə] are the two types of mismatches that were judged as being the most accented. The deletion of /k/ in "*ask*" was considered more accented than vowel epenthesis. Four mismatches of "*ask*" were included in the experiment, namely [ask, æsk, ᵆsk, ᵆsk, ɑsk]. The diacritic marks of [ᵆ] and [ᵆ] indicate that these two pronunciations are sub-phonemic variations of /æ/. In Experiment 1, ratings of the various vowel mismatches in the context of "ask her" did not show any significant difference. In Experiment 2, vowel raising (i.e., [ᵆ]) was rated as more accented than vowel lowering (i.e., [ᵆ]). Ratings of [ask] and [ᵆsk] were not significantly different from ratings of the match stimuli, which seems to indicate that the lowering of /æ/ in "*ask*" is not as accented as vowel raising.

Context 2: "*Please call*"

Mean accentedness ratings are summarized in Figure 5.5. The general trend for the mean ratings of stimuli in the context of "*please call*" as shown in Figure 5.5 is identical to the one found by Experiment 1 (See Figure 4.5).

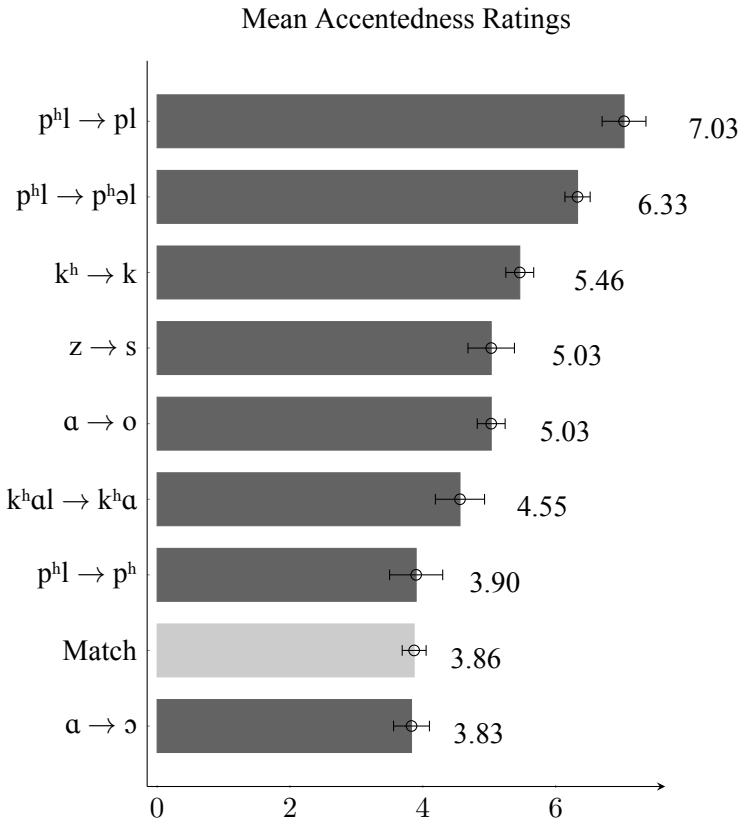

Figure 5.5: Mean Accentedness Ratings of Stimuli in "*Please call*"

Linear mixed-effects regression models were built to achieve pairwise comparisons between the different types of stimuli. Table 5.4 lists the accentedness rankings for stimuli in the context of "*please call*". In the context of "*please call*", consonant mismatches and vowel epenthesis (i.e., [pʰl]→[pʰəl]) were in general rated as more accented than vowel mismatches. The only vowel mismatch being rated as significantly more accented than the match stimuli was [ɑ]→[o]. Rating differences between [ɑ]→[ɔ] and the match stimuli were not significant, probably because /kʰɔl/ is a possible L1 dialectal variation of "*call*." Notably, phrase-initial VOT shortening (i.e., [pʰl]→[pl]) was rated more accented than phrase-medial VOT shortening (i.e., [kʰ]→[k]). Such a finding is consistent with the findings in the Experiment 1.

Table 5.4: Accentedness Ratings for "*please call*"

$p^h l \rightarrow pl$; $p^h l \rightarrow p^h əl$	$\gg$	z→s; ɑ→o	$\gg$	æ→ɑ; Match
$p^h l \rightarrow pl$	$\gg$	$k^h \rightarrow k$	$\gg$	ɑ→ɔ; Match;$p^h l \rightarrow p^h$; $k^h ɑl \rightarrow k^h ɑ$
$p^h l \rightarrow p^h əl$; $k^h \rightarrow k$			$\gg$	ɑ→ɔ; Match; $p^h l \rightarrow p^h$; $k^h ɑl \rightarrow k^h ɑ$
z→s; ɑ→o; $k^h \rightarrow k$ (No significant difference)				

Context 3: "*Six spoons*"

Figure 5.6 demonstrates the mean accentedness ratings for stimuli in the context of "*six spoons*". The general trend for the mean ratings is identical to the one found by Experiment 1.

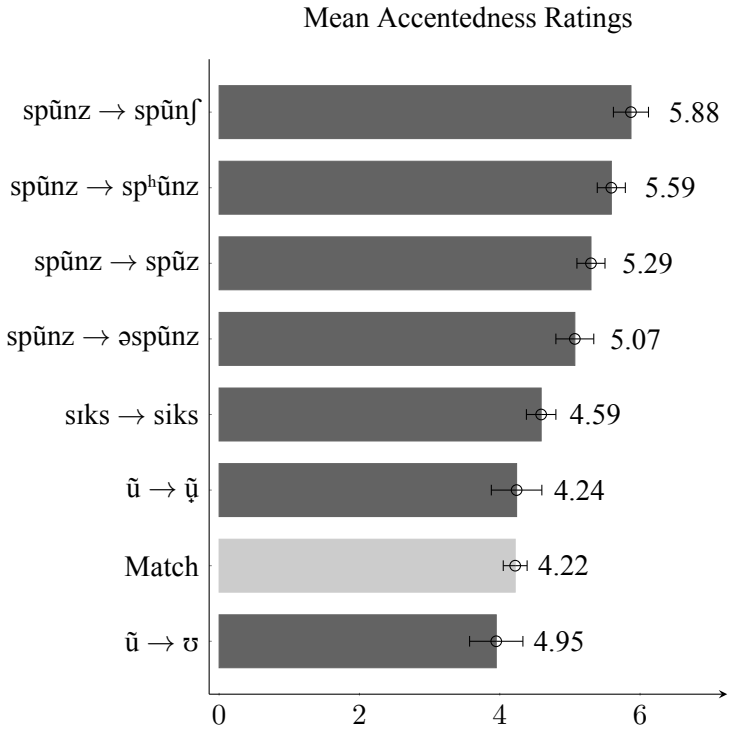

Figure 5.6: Mean Accentedness Ratings of Stimuli in "*Six spoons*"

Table 5.5 lists the accentedness rankings. Only [spũnz]→[spũnʃ],[spũnz]→[spũz] and [spũnz]→[sp^hũnz] received significantly higher accentedness ratings than the match stimuli. In Experiment 1, [spũnz]→[spũnʃ]

received significant higher ratings than [spũnz]→[spʰũnz]. In Experiment 2, the rating difference between the two was not statistically significant. Ratings of prothesis in /sp/ (i.e. [spũnz]→[əspũnz]) and various vowel mismatches were not significantly different from ratings of the match stimuli.

In general, the results show that consonant mismatches were rated as being more accented than vowel mismatches. Analysis on vowel epenthesis in the context "*please call*" shows that vowel anaptyxis (i.e., [pʰl]→[pʰəl]) received significantly higher ratings than ratings of the match stimuli. In the context of "*six spoons*", however, the rating difference between vowel prothesis (i.e., spũnz→əspũnz) and the match stimuli is only marginally significant ($\chi 2 = 3.2$, $p = .06$). These results show that vowel anaptyxis and vowel prothesis might not carry equal weight in accentedness perception.

Table 5.5: Accentedness Ratings for "*Six spoons*"

spũnz→spũnʃ; spũnz→spũz; spũnz→spʰũnz	≫ Match; ũ→ʊ;
spũnz→spũnʃ; spũnz→spũz; spũnz→spʰũnz; spũnz→əspũnz; ũ→ṵ; sɪks→siks	(No significant difference)
ũ→ʊ; Match; ũ→ṵ; sɪks→siks; spũnz→əspũnz	(No significant difference)

Context 4: "*Five thick*"

Figure 5.7 demonstrates the mean ratings of stimuli in the context of "Five Thick." [θ]→[s̪t̪] was rated as the most accented, probably because [θ]→[s̪t̪] involves not only segment replacement but also structural change. This result is consistent with findings in Experiment 1. Interestingly, changing /θɪk/ to /stɪk/ is not a violation of English phonotactics. In fact, the sound sequence /stɪk/ has a higher likelihood to exist in L1 English speech than the sound sequence /θɪk/ (Vitevitch and Luce, 2004). Therefore, raters of the current experiment had probably taken into consideration the lexical outcome (i.e., "*thick*") of the L2 sound sequence (i.e., [s̪t̪ɪk]) while making their accentedness judgment.

Experiment 1 found that vowel mismatches [aɪ]→[ɑɪ] and [aɪ]→[a] were rated as being more

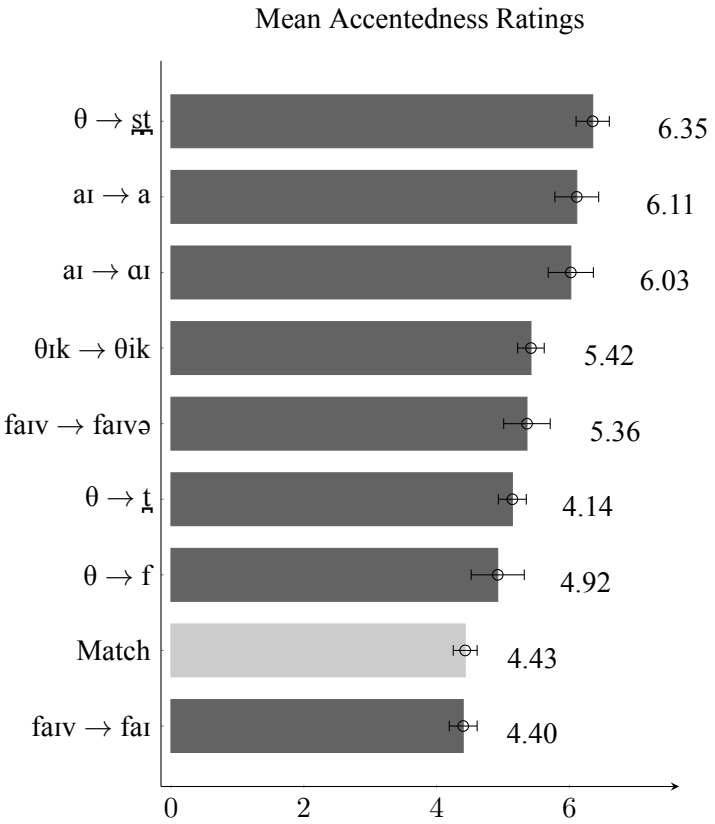

Figure 5.7: Mean Accentedness Ratings of Stimuli in "*Five thick*"

Table 5.6: Accentedness Ratings for "*Five thick*"

θ→st	≫	faɪv→faɪvə; θɪk→θik ≫ faɪv→faɪ; Match
θ→st	≫	faɪv→faɪ; Match; θ→f; θ→t
aɪ→ɑɪ; aɪ→a	≫	faɪv→faɪ; Match; θ→f
aɪ→ɑɪ; aɪ→a; faɪv→faɪvə; θɪk→θik (no significant difference)		

accented than the match stimuli. The same finding was replicated here (Table 5.6). Previous analysis on "*six spoons*" shows that pronouncing "*six*" as [siks] was not significantly more accented than its target pronunciation [sɪks] (i.e., the pronunciation of the match stimuli). Analysis on "*five thick*"

shows that L2 production [θik] was rated as being significantly more accented than the target production [θɪk]. Therefore, the accentedness of vowel tensing (i.e., [ɪ]→[i]) seems to differ depending on phonological contexts.

Context 5: "*Small plastic*"

Figure 5.8 demonstrates the mean ratings of stimuli in the context of "*small plastic*." Consonant mismatches [pʰl]→[pʰɾ],[smɑl]→[smɑ̰l] and [pʰl]→[pʰl̰] were rated as the most accented. Vowel mismatches were relatively less accented.

Table 5.7 shows that /l/-retroflexing and /l/-flapping were rated as the most accented. Ratings of the other types of stimuli were not significantly different from one another. Unlike VOT-shortening on /pl/ in "*please call*", VOT-shortening on /pl/ in the word "*plastic*" was not rated as significantly higher than the match stimuli. These results, again, show that phonological context could have affected accentedness judgment.

5.4.5 Effects of Acoustic Differences

The stimuli of the current study were selected based on the IPA transcriptions, rather than the acoustic information of the segments. IPA transcriptions, in a sense, represent categorical changes in perception, while acoustic information could represent gradient differences in production. Although speech perception for adults is generally categorical, gradient differences in production should not be disregarded. As shown in some previous research, gradient acoustic differences could, to some degree, affect accentedness judgment (McCullough, 2013).

This section discusses the analysis on the effect of gradient acoustic differences on acccentedness perception. As mentioned in the literature review (Chapter 2) and the stimuli selection chapter (Chapter 3), acoustic correlates of a phoneme are multidimensional. One single acoustic measurement might not be representative enough. However, the stimuli used for the current study are limited in their types and tokens, which might not warrant a full investigation of all the relevant acoustic signals. The current study, therefore, investigates only the most commonly used acoustic benchmarks uncovered by previous literature to compare the acoustic difference between an L2 segment and its

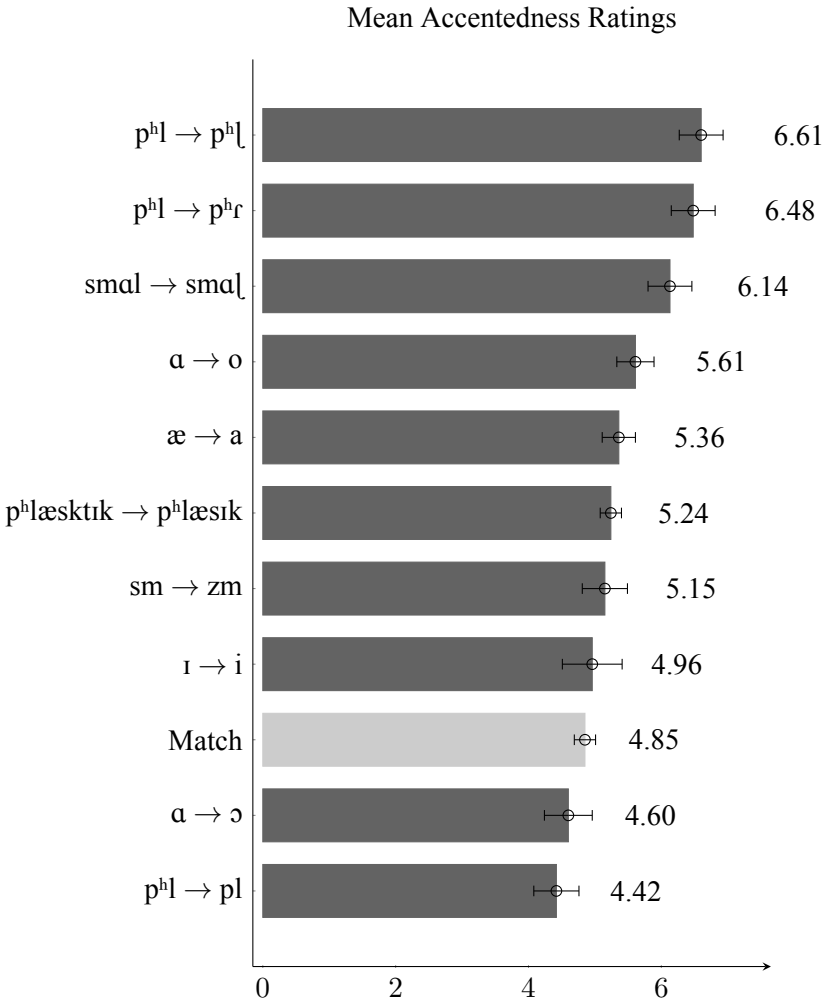

Figure 5.8: Mean Accentedness Ratings of Stimuli in "*small plastic*"

Table 5.7: Accentedness Ratings for "*Small plastic*"

smɑl→smɑɻ; pʰl→pʰɻ; pʰl→pʰɾ	≫	pʰl→pl; ɑ→ɔ; æ→a; Match; ı→i; sm→zm; pʰlæstık→pʰlæsık; ɑ→o;

L1 target. The current study fully acknowledges that some other acoustic signals could also affect how a speech sound is perceived.

As discussed extensively in previous research, phonological contexts have crucial impacts on acoustic signals of segments. It is therefore essential for the current study to control for phonological context. Due to the limitation of the current research design, acoustic analysis could only focuse on four types of patterns in L2 speech, namely VOT-shortening, /θ/-related stimuli, /æ/-related stimuli, and stimuli with vowel epenthesis. For plosive segments with VOT-related variations, the durations of the VOTs were measured and subsequently compared to mean L1 VOT duration values. Absolute z-scores were calculated to approximate the acoustic distance between an L2 plosive and its L1 target.

For fricatives, Center of Gravity values (COG), which represent place of articulation and voicing, were calculated and compared to mean L1 COG values. Absolute z-scores were calculated to approximate the acoustic differences between L2 fricatives and their L1 targets. For manner of articulation, noise ratio was used to approximate L2 pronunciation [t]s to their L1 target production [θ]. The definition of noise ratio is the same as described in Chapter 3.

For vowel /æ/, F1 and F2 values were calculated to approximate tongue position. The Euclidean distances between L1 and L2 F1/F2 values were calculated to represent the acoustic difference between L1 and L2 vowels. Male and female speech samples were compared separately. The details of the method and calculation are as described in Chapter 3.

Both Experiment 1 and Experiment 2 show that the prothesis of /sp/ in the context of "*six spoons*" (i.e., [sp]→[əsp]) was not judged as being more accented than the match stimuli, while vowel anaptyxis of /pl/ in the context of "*please call*" (i.e., [pʰl]→[pʰəl]) and vowel paragoge in contexts of "*five thick*" (i.e., [faɪv]→[faɪvə]) and "*ask her*" (i.e., [æsk]→[æskə]) were rated as being significantly more accented than their respective match stimuli. This section introduces an investigation on whether epenthetic vowel duration had an effect on accentedness judgment.

VOT-Shortening

For the investigation of gradient differences of VOT, the context "*please call*" was chosen. Three stimuli with VOT-shortening and five match stimuli were chosen for the analysis. Table 5.8 illustrates the three stimuli with VOT-shortening. For example, the first row shows the specific stimulus is in the context of "*please call*", the type of mismatch is the shortening of VOT on the /k/ of "*call*". The VOT duration of this specific [k] is 33 ms and the VOT duration for the [p] in "*please*" is 40 ms.

Table 5.8: L2 Stimuli with VOT-related Mismatches

Context	Type of Stimuli	VOT (ms)
please call	VOT-shortening on /k/	[p] = 40; [k] = 33
please call	VOT-shortening on /k/	[p] = 32; [k] = 21
please call	VOT-shortening on /p/	[p] = 10, [k] = 62
please call	match	[p] = 48; [k] = 51
please call	match	[p] = 63; [k] = 55
please call	match	[p] = 54; [k] = 62
please call	match	[p] = 55; [k] = 63
please call	match	[p] = 50; [k] = 42

VOT durations of the [p]s and [k]s for all eight stimuli were measured. Absolute z-scores were calculated by comparing VOT length values of the stimuli to L1 VOT values. As introduced in the stimuli selection chapter (Chapter 3), mean L1 VOT length and its standard deviation were calculated based on 50 L1 American English speakers' productions. The mean L1 VOT duration for the /p/ in "*please*" is 62.5 ms (SD=18.60). The mean L1 VOT duration for the /k/ in "*call*" is 52.78 ms (SD=13.71).

Mixed effects linear regression models were used to investigate the effect of acoustic distances on accentedness. Ratings were the dependent variable. Absolute Z-scores for [p]s and [k]s were entered as a fixed effect. Type of stimuli (i.e., match vs. VOT-shortening) was another fixed effect. Type of segments (i.e., /p/ vs. /k/) was the third fixed effect. The two-way and three-way interactions of the aforementioned three fixed effects were also entered as fixed effects. Raters were entered as

a random effect with "type of stimuli" as the random slope. The stimuli were entered as another random effect.

Model comparisons using the likelihood ratio test revealed that the type of stimuli (i.e., match vs. VOT-shortening) as the only factor that significantly affected model fit ($\chi 2 = 7.01$, $p < .05$). Absolute Z-scores and type of segments did not contribute to model fit significantly. These results indicate that VOT durations of /p, k/s in the context of "*please call*" did not significantly affect accentedness ratings.

Fricatives

Place of Articulation The phrase "*five thick*" was chosen to investigate the pronunciation of /θ/. Four mismatch and five match stimuli were chosen for the following analysis. Three of the four mismatch stimuli involve pronouncing /θ/ as [t̪] or [t]; One involves pronouncing /θ/ as [f]. The /θ/s of the match stimuli were all transcribed with [θ]. As introduced in Chapter 3, Center of Gravity (COG) could represent place of articulation and voicing differences between fricatives. At the same time, plosives in general have a lower COG than fricatives. COG was therefore chosen as a benchmark acoustic measurement to represent gradient place difference between the L1 and L2 segments.

The L1 COG values were calculated based on 50 L1 American English speakers' productions selected from the SAA. L2 COGs were compared to L1 COGs according to the gender of the speaker. Absolute z-scores were computed to represent how much the L2 segments differ from the L1 means. L1 mean COGs were calculated using 50 L1 American English speakers' production of "*five thick*". The mean COG for male L1 speakers' production of /θ/ is 64.37 semitones (SD=11.40). The mean COG of female L1 speakers is 64.86 semitones (SD=9.71). Details of the data were introduced in Chapter 3 and re-listed below in Table 5.9.

Linear mixed-effects regression models were employed to investigate the effect of gradient COG values on accentedness ratings. Ratings were used as the dependent variable. Absolute z-score values, the type of stimuli (i.e., match vs. mismatch) and the interaction between the two were entered as fixed effects. Raters were entered as a random effect with "type of stimuli" as the random slope. Stimuli were entered as another random effect. Model comparisons using the likelihood ratio

Table 5.9: L2 COGs (Semitone)

Context	Type of Stimuli	L2 COGs
five thick	mismatch (/θ/→[f])	73.77
five thick	mismatch (/θ/→[t̪])	59.63
five thick	mismatch (/θ/→[t̪])	63.27
five thick	mismatch (/θ/→[t̪])	42.63
five thick	match	70.53
five thick	match	61.7
five thick	match	80.6
five thick	match	55.53
five thick	match	78.1

tests revealed that "type of stimuli" is the only factor that contributed significantly to model fit ($\chi 2$ = 4.44, $p < .05$). Absolute z-scores and the interaction between absolute z-scores and "type of stimuli" did not contribute significantly to model fit. In other words, no evidence could support the claim that gradient COG differences affected accentedness judgments.

Manner of Articulation As discussed in Chapter 3, the difference between /t/ and /θ/ lies in their respective manner of articulation, which was not captured by the COG measurements. The manner difference between English fricatives and plosives could potentially be approximated by the duration of frication noise. As Jongman (1989) discovered, the shorter the duration of friction noise, the more likely a fricative is perceived as a plosive. Noise ratio were calculated to approximate the gradient acoustic difference between /θ/s and the /t/s. Noise ratio was defined as the ratio of noise duration over the duration of the whole word (Jongman, Wayland, and Wong, 2000). Noise duration for /t/s was defined as the interval from the beginning of the bursts to the beginning of the following vowel. Noise duration for /θ/s was defined as the duration of the frication noise of the /θ/s. The mean noise ratio for male L1 speakers is 0.25 (SD=0.06), while the mean noise ratio for female L1 speakers is 0.20 (SD=0.08). Absolute z-scores were computed to represent how much the L2 segments differ from the L1 means. Details of the measurement and calculation are introduced in Chapter 3. Table 5.10 lists the details of the data.

Table 5.10: L2 Noise Ratios

Context	Type of Stimuli	Noise Ratio
five thick	mismatch (/θ/→[t̪])	0.15
five thick	mismatch (/θ/→[t̪])	0.17
five thick	mismatch (/θ/→[t̪])	0.16
five thick	match	0.21
five thick	match	0.37
five thick	match	0.17
five thick	match	0.32
five thick	match	0.18

Linear mixed-effects regression models were employed to investigate the effect of noise ratio on accentedness ratings. Ratings were used as the dependent variable. Absolute z-scores for noise ratio, Type of stimuli (i.e., match vs. mismatch) and the interaction between the two were entered as fixed effects. Raters were entered as a random effect with type of stimuli as the random slope. Stimuli were entered as another random effect. Model comparison using likelihood ratio tests revealed that "type of stimuli" is the only factor that contributed significantly to model fit ($\chi 2 = 5.04$, $p < .05$). Absolute z-scores for noise ratio and the interaction between absolute z-scores and "type of stimuli" did not contribute significantly to model fit. In other words, no evidence could support the claim that gradient differences in noise ratio affected accentedness judgment.

Vowels

The context "*ask her*" was chosen to investigate the effect of spectral information on vowel accentedness. Ten stimuli were chosen for the analysis. Five of them involve the raising, lowering, and backing of the phoneme /æ/ in "*ask*" as indicated by their respective IPA transcriptions. Another five stimuli are L2 productions containing no segmental or structural mismatches (i.e., the match stimuli). 50 L1 American English speech samples were selected to extract L1 F1 and F2 values for the /æ/ in "*ask*". L1 Mean F1 and F2 values were calculated for comparisons. Acoustic distances between L2 segments and L1 productions were defined as the Euclidean distances in the F1-F2 space,.

The F1 and F2 values of the ten L2 productions of /æ/ were compared to L1 mean F1 and F2 values using Equation 5.1. The "distances" column in Table 5.11 shows the Euclidean distance from each L2 production of the /æ/ to L1 mean values.

$$Distance = \sqrt{(L1\ Mean\ F1 - Stimulus\ F1)^2 + (L1 Mean\ F2 - Stimulus\ F2)^2} \quad (5.1)$$

Table 5.11: Euclidean Distances between L2 Vowels and L1 Means

Context	Type of Stimuli	Distances
ask her	mismatch (/æ/→[ɑ])	7.30
ask her	mismatch (/æ/→[a̟])	6.57
ask her	mismatch (/æ/→[a])	5.63
ask her	match	4.82
ask her	mismatch (/æ/→[a])	4.55
ask her	match	4.32
ask her	mismatch (/æ/→[a̟])	2.44
ask her	match	2.12
ask her	match	1.55

Linear mixed-effects regression models were again employed to investigate whether acoustic distance affects accentedness ratings. Ratings were dependent variables. Euclidean distances, type of stimuli (i.e., match vs. mismatch), and the interaction between the two were entered as fixed effects. Raters were a random effect with "type of stimuli" as the random slope. Stimuli were another random effect. Model comparisons using likelihood ratio tests revealed no significant effects to model fit. That is, no evidence was found to support the claim that gradient acoustic distance effected accentedness perception. Unlike findings for obstruent consonants, "type of stimuli" was not found to be a significant contributing factor. That is, L2 productions [ask, ɑsk, a̟sk, a̟sk] as a whole were not judged to be more accented than /æsk/.

Vowel Epenthesis

Experiment 1 and 2 investigated the effect of vowel epenthesis on accentedness judgment. Three types of vowel epenthesis were included, namely prothesis of an s-cluster (i.e., [sp]→[əsp]), anaptyxis of /pl/ (i.e., [pʰl]→[pʰəl]) and paragoge (i.e., [faɪv]→[faɪvə], [æsk]→[æskə]). Durations of the epenthetic vowels could have affected accentedness judgment. Therefore, durations of the epenthetic vowels were measured. To account for speech rate, epenthetic vowel ratios were calculated by taking the ratio of the epenthetic vowel duration over the duration of the whole word. Table 5.12 below lists the details of the data.

Table 5.12: Duration of the epenthetic vowels

Mismatches	Contexts	Ratio	Type of Stimuli
pʰl→pʰəl	*please call*	0.11	anaptyxis
pʰl→pʰəl	*please call*	0.09	anaptyxis
pʰl→pʰəl	*please call*	0.06	anaptyxis
faɪv→faɪvə	*five thick*	0.13	paragoge
æsk→æskə	*ask her*	0.19	paragoge
æsk→æskə	*ask her*	0.34	paragoge
spũnz→əspũnz	*six spoons*	0.07	prothesis
spũnz→əspũnz	*six spoons*	0.03	prothesis

There were three tokens for [pʰl]→[pʰəl] in the context of "*please call*", one token for [faɪv]→[faɪvə] in the context of "*five thick*", two tokens for [æsk]→[æskə] in the context of "*ask her*", and two tokens [spũnz]→[əspũnz] in the context of "*six spoons*". The "Ratio" column listed the ratio of epenthetic vowel duration over the duration of the whole word. Data in Table 5.12 shows that the durations of the prothetic vowels in context "*six spoons*" were generally shorter than other types of epenthetic vowels, which might explain why [əspũnz] was judged as relatively less accented.

In order to investigate the effect of epenthetic vowel duration on accentedness perception, the eight stimuli were grouped into three types, as shown in the "Type of Stimuli" column in Table 5.12. Since the three types of epenthesis happened in four contexts, ratings of the match stimuli in these four contexts were also included. Since the match stimuli do not contain vowel epenthesis, the

epenthetic vowel ratios for these match stimuli were defined as 0.

Linear mixed-effects models were used for the analysis. Ratings were the dependent variable. Epenthetic vowel ratios, type of stimuli (i.e., match vs. anaptyxis vs. paragoge vs. prothesis) and the interaction between the two were used as fixed effects. Raters were a random effect with type of stimuli as the random slope. Stimuli were another random effect. To achieve pairwise comparisons of the four type of stimuli, the "type of stimuli" variable was contrast-coded using Helmert contrasts. The first contrast compares anaptyxis with paragoge, the second contrast compares prothesis with the other two types of vowel epenthesis, and the third contrast compares the three types of vowel epenthesis with the match stimuli.

Model comparisons using likelihood ratio tests show that the third contrast contributed significantly to model fit ($\chi 2 = 11.40$, $p < .01$), showing that the match stimuli in the four contexts received significantly lower ratings than the three types of vowel epenthesis. The contribution of epenthetic vowel ratio to model fit was not significant ($\chi 2 = 0.53$, $p = .47$). In other words, no evidence was found to support the claim that the duration of epenthetic vowels affected raters' accentedness judgment.

5.4.6 Summary

Results of Experiment 2 generally agree with results of Experiment 1. Consonant and syllable mismatches were judged to be more accented than vowel mismatches, which were judged as being more accented than the match stimuli. Analysis on individual mismatches demonstrates that not all mismatches were weighted the same in acccentedness perception. The same type of mismatch (e.g., VOT-shortening) could be weighted differently depending on where it occurred. Trial number positively correlated with ratings, showing that raters became less lenient as the experiment progressed. The same pattern was also observed in Experiment 1.

Experiment 2 also investigated the possible effect of gradient acoustic differences on accentedness. Since the data of the current study are limited in their types and tokens, no conclusive results can be reported at this time. It is highly possible that sub-phonemic acoustic differences were not a deciding factor for accentedness judgments, especially not for the judgment on obstruent consonants.

5.5 Discussion

Like Experiment 1, Experiment 2 shows that different types of mismatches did not carry the same weight in the raters' accentedness judgment. Three observations can be reached from the results of Experiment 1 and Experiment 2.

First, mismatches that could be considered L1 dialectal variations were rated as less accented. For example, pronouncing "*thick*" as [tɪk] or [fɪk] was not judged to be more accented than pronouncing "*thick*" as [θɪk] (i.e., the match stimuli). Substituting /θ/ with /f/ or /t/ is a prominent feature in many varieties of African American Vernacular English (Green, 2002). In addition, /θ/ is sometimes realized as /t/ or /tθ/ in New York City, Philadelphia and other northern cities in the U.S. (Gordon, 2008). Substituting /θ/ with /f/ or /t/ is also found in varieties of British English (Altendorf and Watt, 2008), Australian English (Horvath, 2008) and Newfoundland English (Clarke, 2008). Raters who are familiar with L1 dialectal variations of /θ/ would probably consider [tɪk] or [fɪk] less foreign accented. On the other hand, pronouncing "spoons" as [spũnʃ] is not an L1 dialectal variation. [spũnʃ] was indeed judged to be very accented.

Raters of the current study assigned lower accentedness ratings to [kʰɔl] and [smɔl] (i.e., less accented), probably because [kʰɔl] and [smɔl] are possible L1 variations for "*call*" and "small". Over 70% of the 100 L1 American English speakers surveyed by the current study pronounced "*call*" and "small" as [kʰɑl] and [smɑl], which could be a result of the COT-CAUGHT merger found in many varieties of American English (Labov, Ash, and Boberg, 2005). The COT-CAUGHT merger refers to the merging of vowel phonemes [ɑ] and [ɔ] into the same phoneme [ɑ]. L1 English varieties that do not have the COT-CAUGHT merger could pronounce "*call*" and "*small*" as [kʰɔl] and [smɔl]. The lower accentedness ratings of [kʰɔl] could have resulted from raters' familiar with L1 English varieties that do not have the COT-CAUGHT merger.

Other than the COT-CAUGHT merger, the current study also included a case of off-glide deletion (i.e., [faɪv]→[fav]) that resembles the off-glide deletion phenomenon observed in many varieties of Southern American English (Labov, Ash, and Boberg, 2005). Off-glide deletion of [aɪ] in Southern American English deletes the glide [ɪ] and lengthens the [a]. The word "*five*" in Southern American

English could be pronounced as [fa:v], but not [fav]. Therefore, [fav] is, strictly speaking, not native-like. Raters of Experiments 1 and 2 indeed assigned higher accentedness ratings to [fav] (i.e., more accented). The difference between [fa:v] and [fav] lies in the duration of the [a]. The current study did not measure vowel duration. Further study is needed to investigate the effect of vowel duration on accentedness perception.

The second observation one could reach from results of Experiments 1 and 2 is that phonological context affects perceptual accentedness. For example, VOT shortening on /pl/ clusters was judged as very accented phrase-initially, but less accented phrase-medially. In the context "*please call*," pronouncing the word "*please*" as [pliz] was judged as being more accented than pronouncing it as [pʰliz]. In the context "*small plastic*", pronouncing the word "*plastic*" as [plæstɪk] was not significantly more accented than pronouncing it as [pʰlæstɪk]. This result agrees with the L1 speech data in the SAA. Among the 100 L1 speakers of American English surveyed by the current study, only 11 of them pronounced the "*p*" in "*please*" (in context "*please call*") as an unaspirated [p]; while 30 of them pronounced the "*p*" in "*plastic*" (in context "*small plastic*") as an unaspirated [p]. Since VOT-shortening is more likely to happen phrase-medially than phrase-initially in L1 speech, it is therefore understandable that phrase-medial VOT-shortening was judged as less accented.

The accentedness of vowel tensing was also different depending on the phonological context. For example, pronouncing "*thick*" as [θik] was judged as more accented than the L1 target production [θɪk]. However, pronouncing "*six*" as [siks] was not judged as more accented than its target L1 production [sɪks]. This phenomenon could be attributed to English phonotactics. Although [siks] is not a native-like production of "*six*", the pronunciation itself exists in English (e.g., "seeks"). Sound sequence [θik], on the other hand, rarely occurs in English.[2] It is, therefore, possible that the sound sequence [θik] sounded more accented to the raters. Alternatively, the degree of vowel tensing could have affected accentedness perception. It is possible that the [i] in [θik] is more tensed than the [i] in [siks]. Stimuli of the current study are limited in their types and tokens, which cannot warrant a detailed acoustic investigation on the tenseness of the vowels. Future study is needed to uncover the relationship between the degree of vowel tenseness and accentedness perception.

[2]None of the 133031 English words collected by the CMU English pronunciation dictionary contains sound sequence [θik] (Weide, 1998)

For stimuli with syllable mismatches, Experiments 1 and 2 predicted that stimuli with coda consonant deletion should be less accented than stimuli with vowel epenthesis, because coda consonant deletion is allowed in L1 English speech. Results of Experiment 1 and 2 show that pronouncing "*five*" as [faɪ] was not judged as significantly more accented than pronouncing "*five*" as [faɪv]. On the other hand, stimuli with vowel epenthesis such as [pʰl]→[pʰəl] and [æsk]→[æskə] was judged to be relatively more accented. Although coda consonant deletion often happens in L1 English speech, pronouncing "*ask*" as [æs] was judged to be very accented. The specific phonological context for the "*ask*" is "*ask her*", which does not usually allow [k]-dropping in L1 speech. The relatively higher accentedness ratings of [æs], therefore, indicated that the raters have taken into consideration the phonological context when making accentedness judgments.

The third observation is that ratings of both Experiments 1 and 2 increased over time. Since raters of Experiment 1 were not aware of the intended meaning of each stimulus, it is possible that intelligibility could have played a role in raters' accentedness judgment. That is, the raters could have made their judgment based on whether they could understand the stimuli. For example, pronouncing [sɪks] as [siks] might be considered native-like when the raters did not know the correct lexical outcome at the earlier stage of the experiment. [siks] would be considered accented when raters became aware of the correct lexical outcome later in the experiment. Experiment 2 controlled for the impact of intelligibility by explicitly telling raters what they were going to hear before they were exposed to the audio stimuli. Therefore, the effect of intelligibility might not be sufficient to explain the increasing of ratings over time.

The reason for the increase in ratings could have resulted from the presentation of the stimuli. As stated earlier, a block randomization strategy was implemented to counterbalance order effects. The presentation of stimuli was organized in the way that a native-like stimulus (i.e., the match stimuli) occurs in every five trials. Participants' exposure to native-like stimuli, thus, gradually increased as the experiment progressed. The increased exposure to native-like stimuli could have made the mismatches contained in the stimuli perceptually more salient, which would lead to higher accentedness ratings. Such a finding is consistent with Flege and Fletcher (1992), who similarly reported that the portion of native-like stimuli included in an experiment affects raters' judgments

on the accentedness of L2 stimuli.

Although Experiments 1 and 2 produced similar results, the two experiments do differ in their respective research designs. A training phase was added to Experiment 2 to familiarize raters with the range of accentedness contained in the experiment. The training phase has, to some degree, facilitated the testing phase. As shown in Figure 5.3, the ratings of Experiment 2 were more consistent and less variable than ratings of Experiment 1. However, the ten stimuli used in the training phase could also have potentially biased raters' judgments during the testing session. The ten training session stimuli were selected based on their accentedness ratings obtained by the Experiment 1. By including such a training phase, raters of Experiment 2 were therefore encouraged to behave more like raters of Experiment 1.

Experiment 2 also controlled for intelligibility by explicitly informing raters of the lexical outcomes of the stimuli. It should be noted that informing raters of the lexical outcomes is not necessarily a better design. In real life situations, it is not always the case that a naïve L1 listener is aware of the intended meaning of an L2 pronunciation. Therefore, the two experiments represented two possible scenarios in real life. Experiment 1 represents the situation when the listeners are not aware of the intended meaning. Experiment 2 represents the situation when the listeners are aware of the intended meaning. Analysis in the current chapter shows that accentedness judgment becomes harsher when the intended meaning is known. However, miscommunication could arise when the intended meaning of an L2 speech is unknown. The current study focuses only on the accentedness, rather than the intelligibility of an L2 speech. Further research is needed to investigate the relationship between accentedness and intelligibility, and their respective sociolinguistic consequences.

Several previous studies had attempted to separately investigate accentedness, nativelikeness, intelligibility and comprehensibility (Derwing and Munro, 1997; Kennedy and Trofimovich, 2008; McCullough, 2013). It is undeniable that these terms refer to different aspects of human speech. The question is if a naïve L1 listener would be aware of the subtle differences between these aspects and make his or her accentedness judgment accordingly. In lab-settings, clear definitions could be given. Experimenters could then force raters to abide by these definitions. Many previous studies have indeed implemented such a practice.

The current study, although it provided a definition of "foreign accent" in Chapter 1, did not define this term for the raters. In fact, no definition of "accentedness" or "foreign accent" was given at any point of the two experiments. The experiments simply instructed raters to rate how "foreign accented" the stimuli sounded. The purpose for such design was to emulate real life situations when no rules and definitions are given. A naïve L1 listener would have to rely on his or her own understanding of accentedness to make the judgment.

The raters recruited by the current study undoubtedly have their own definitions of foreign accentedness and these definitions could be very different from one another. This situation is very much like what might happen in real life when a naïve listener has to make the judgment based on his or her intuition. It is the responsibility of the researchers to understand naïve listeners' intuition, rather than to impose academic guidelines onto them for the purpose of experimental control.

From the perspective of phonology and phonetics, one possible component of naïve L1 listeners' intuition is undoubtedly their L1 phonological/phonetic knowledge. An L2 speech pattern would be considered accented once it deviates from what is considered native-speaker norms. Or in Major (2012, p.1)'s words, foreign accent is "*a pronunciation deviating from what a native speaker expects another native speaker to sound like.*" If Major's definition is accepted, then the issues at hand are to investigate what an L1 speaker should sound like and to formulate a method for the measurement of phonological/phonetic deviation. The next chapter will address these two issues and further attempts to uncover the possible mechanisms underlying the perception of foreign accentedness.

Chapter 6: Experiment 3

6.1 Introduction

Results of Experiments 1 and 2 demonstrates that phonetic, phonological, and lexical information was being taken into consideration in raters' accentedness judgments. For example, /æ/-raising in "*ask*" was found by Experiment 2 to be more accented than /æ/-lowering, showing that sub-phonemic vowel information could potentially be perceived and rated differently. VOT-shortening was rated as more accented phrase-initially than phrase-medially, showing that phonological context plays a role in accentedness judgments.

Previous literature often defines the "accent" of a pronunciation in terms of phonetics and phonology, without regard to the lexical aspect of the pronunciation (Wells, 1982). Although accentedness is not equivalent to intelligibility or comprehensibility, naïve raters recruited by the current study seem to have utilized lexical information in their accentedness judgment. More specifically, non-native (L2) stimuli were rated as being more accented once their intended meanings were known. Therefore, the current study hypothesizes that lexical information of the stimuli together with English phonetic and phonological grammar (L1 knowledge) could have affected raters' accentedness judgment.

To investigate this hypothesis, Experiment 3 was conducted. This chapter discusses Experiment 3, which computationally models raters' knowledge of English phonetics and phonology with regard to the five phonological contexts (i.e., "*ask her*," "*please call*," "*five thick*," "*small plastic*," "*six spoons*"). Lexical information of the five contexts was also taken into consideration. The 100 L2 stimuli used by Experiments 1 and 2 were evaluated by the model to generate dissimilarity scores, approximating the degree of dissimilarity between the L2 stimuli and their corresponding L1 productions. Ratings from Experiment 2 were further compared against the dissimilarity scores to investigate whether and how raters' L1 knowledge affected their accentedness judgment.

6.2 Dissimilarity Measurements

Previous research on foreign-accented speech has often investigates dissimilarities between L1 and L2 speech based on their respective benchmark acoustic signals (e.g., VOT, vowel formant frequencies, etc.) (e.g., McCullough, 2013). Experiment 2 (Chapter 5) of the current study attempted comparing acoustic signals of a selection of L1 and L2 speech samples. The results, however, are not entirely conclusive due to the limited types and tokens of mismatches portrayed by the 100 audio stimuli of the current study.

Instead of analyzing acoustic signals, some studies in the field of dialectology measured the degree of dissimilarity between dialects/languages in terms of "phonetic distances" (e.g., Nerbonne et al., 1996). These studies usually obtained "phonetic distances" by instructing speakers of different dialects/languages to read the same list of words (e.g., Nerbonne et al., 1996). Their pronunciations of the words were then transcribed into IPA symbols. These transcriptions for different speakers were compared against each other using alignment algorithms such as the Levenshtein distance measurement (Nerbonne et al., 1996; Wieling et al., 2014a) or the Dynamic Time Warping approach (e.g., Johnson, 2004a). The alignment costs were usually termed "phonetic distances" and were used to approximate the degree of dissimilarity between dialects/languages. The calculation of the phonetic distance was not based on acoustic measurements, but IPA transcriptions. The dissimilarity measurement was termed "phonetic distance" or "pronunciation distance, " because it was based on pronunciations of the same list of words, rather than morphological or syntactic differences between dialects/languages.

It is worth noting that "phonetics" as a field of study concerns the production and perception of both segmental and prosodic information of human speech. "Phonetic distance" in the aforementioned studies was only a measurement of perceptual differences between speech samples, because it resulted from comparisons between IPA transcriptions, which mostly rely on transcribers' perception of speech sounds. In addition, the calculation of "phonetic distance" was based on phonetic segments rather than prosody. In other words, "phonetic distance" is a measurement in segmental dimensions. The measurement of phonetic distance has been adopted to compare L1 English speech with foreign-accented English speech using either the Levenshtein distance (e.g., Schaden, 2006;

Wieling et al., 2014a) or a Dynamic Time Warping approach (e.g., Shen et al., 2013). However, these methods are not cognitively grounded, and merely serve as conveniences for computation. No arguments have been offered to support the claim that algorithms of these methods are representative of human perception. We therefore need a dissimilarity measurement that could potentially reflect human speech perception.

6.3 The Naïve Discriminative Learning Model

Although debates still exist, most research on speech perception considers statistical learning one of the mechanisms underlying the acquisition and processing of language (e.g., Romberg and Saffran, 2010). That is, learners use distributional properties of linguistic input to discover patterns. Several computational models within the framework of statistical learning have been proposed to approximate English phonetic and phonological grammar based on either co-occurrences of distinctive features (Hayes and Wilson, 2008) or co-occurrences of phonetic segments (Vitevitch and Luce, 2004; Wieling et al., 2014b).

The Naïve Discriminative Learning (NDL) approach, as implemented by Baayen (2011), is a method within the framework of statistical learning. It attempts to find the statistical relationship between certain phonetic cues (e.g., [æsk]) and lexical outcomes (e.g., "*ask*"). The NDL approach is based on the Rescorla-Wagner learning theory (Rescorla and Wagner, 1972), which assumes that learners attempt to predict an outcome based on available cues. The association strength from a set of cues to a certain outcome increases if the set of cues often associates with the outcome. Alternatively, if a set of cues rarely associates with a certain outcome, the association strength from the set of cues to the outcome is weaker. Rescorla and Wagner (1972) formulated such process into Equations 6.1 and 6.2.

$$V_i^{t+1} = V_i^t + \Delta V_i^t \tag{6.1}$$

Equation 6.1 defines the association strength at time *t + 1* (i.e., V_i^{t+1}) as the previous association strength at time *t* (i.e., V_i^t) modified by some change in association strength ΔV_i^t. The change of

association strength ΔV_i^t is further defined in Equation 6.2.

$$\Delta V_i^t = \begin{cases} 0 & if_{ABSENT}(C_i, t) \\ \alpha_i \beta_1 (\lambda - \sum_{present(C_j,t)} V_j) & if_{PRESENT}(C_j, t) \& _{PRESENT}(O, t), \\ \alpha_i \beta_2 (\lambda - \sum_{present(C_j,t)} V_j) & if_{PRESENT}(C_j, t) \& _{ABSENT}(O, t), \end{cases} \tag{6.2}$$

Equation 6.2 defines the change in association strength ΔV_i^t in terms of the relationship between cues (C) and a certain outcome (O). If a cue is absent at time *t*, then the association strength is unchanged. If both the cue and the outcome are present, then the association strength increases. If the cue is present in the absence of the outcome, then the association strength decreases.

By way of example, consider the top five most likely L1 pronunciations for the word "*ask*," as shown in Table 6.1. Data in Table 6.1 were extracted from productions of 100 L1 speakers of American English in the SAA. For illustration, only the top five most likely pronunciations of "*ask*" are listed.

Table 6.1: The Top Five most likely L1 Pronunciations of "*Ask*"

Outcome	Cues	Frequencies
ask	æsk	77
ask	æ̟sk	9
ask	æ̟sk	3
ask	æks	2
ask	æ̃sk	2

For an infant exposed to these possible pronunciations of "*ask*," Rescorla-Wagner's theory would predict that the association strength from phonetic cues [æsk] to the lexical outcome "*ask*" increases every time the infant hears "*ask*" being pronounced as [æsk]. At the same time, the association strength from other cues (e.g., [æks]) to the lexical outcome "*ask*" would be less. The frequency

data in Table 6.1 indicate that the association strength from [æsk] to "*ask*" is the strongest, because [æsk] associates more often with "ask."

For infants whose L1 system is not stable, the association strength between the outcome and its possible cues (e.g., pronunciations) is constantly updating, depending on the type and the amount of language input. For adults whose L1 system is relatively stable, further L1 language input will no longer substantially alter the association strengths from cues to outcomes (Chamorro, Sorace, and Sturt, 2016). That is, one could define association strengths for adults by assuming that the association strength at time *t+1* is always the same as the association strength at time *t*. Based on Rescorla-Wagner's learning theory, Danks (2003) derived a set of equations defining stable stage association strengths from cues to outcomes. Baayen (2011) further applied Rescorla and Wagner (1972) and Danks' (2003) equations to the analysis of association strength from linguistic cues to lexical outcomes. Details of the computation can be found in Baayen (2011).

Wieling et al. (2014b) applied Baayen (2011)'s NDL approach in accentedness reasearch. Their model mainly concerns two types of frequencies: (1) frequency of a lexical outcome in L1 speech (e.g. how often the word "*ask*" is used in English); (2) trigram frequencies of segment sequences that occur with the outcome (e.g. how often a pronunciation of "*ask*" contains sound sequence [æsk]). To examine the validity of the NDL approach, Wieling et al. (2014b) first obtained IPA transcriptions from 115 L1 American English speech samples and 280 L2 speech samples in the SAA. 58 L1 speech samples were randomly selected to construct a "native speaker model." Each possible pronunciation (e.g. [æks, ɑsk, æsk, æs]) for its lexical outcome (e.g. "*ask*") was assigned an "association strength" based on two types of frequencies mentioned above, representing the probability of an L1 American English speaker producing the word using that specific pronunciation. The "native speaker model" was therefore a matrix of association strengths, mapping pronunciations to lexical outcomes.

The rest of the 57 L2 speech samples were evaluated using the constructed native speaker model to generate a matrix of association strengths for speech productions of an "average" L1 American English speaker. 280 L2 speakers' productions were also evaluated by the native speaker model to generate association strengths for speech productions of each L2 speaker. Association strengths for each of the L2 productions were then compared to association strengths of the average L1 speaker's

production. The model then generated a "pronunciation distance," representing how much each L2 production differs from an average L1 American English speaker's production. The larger the pronunciation distance, the more different a given L2 speech production would be from an average L1 American English speaker's production. Presumably, the larger the pronunciation distance, the less native-like a given L2 speech sample is.

Wieling et al. (2014b) conducted a perception study to further evaluate his NDL model. The perception experiment obtained native-likeness ratings of the L1 and L2 speech samples. Results show that the pronunciation distances (i.e., non-nativelikeness) strongly correlate with native-likeness ratings provided by 1143 L1 American English listeners (r = -0.72), lending support to the validity of the NDL approach.

Although the NDL-based pronunciation distances correlate well with empirical perception ratings, a question remains as to whether the NDL approach truly represents accentedness perception. Wieling et al. (2014b)'s pronunciation distances were generated based on speech samples each consisting of 69 words with more than 200 segments (i.e., the "Stella" passage in the SAA). It is doubtful that a listener could remember all the segments and further mentally calculate their phonological and phonetic departures from an average L1 production.

Previous empirical research on accentedness detection has shown that L2 accent could be detected within a very short amount of time (Flege, 1984; Park, 2013). It is possible that accentedness judgment could have been formulated after hearing just the first few words or sentences. On the other hand, the storage of short-term memory is limited and memory itself is subject to decay over time (Berman, Jonides, and Lewis, 2009; Miller, 1956). It is, thus, equally possible that accentedness judgment was based on the last few words or sentences heard. No matter what the scenario is, it is unlikely that acccentedness ratings obtained by Wieling et al. (2014b) were based on the whole 69-word-long utterance.

In summary, the NDL approach is in the framework of statistical learning. More specifically, it assumes that humans are capable of generalizing patterns in their native language and use these patterns to evaluate the accentedness of an L2 speech utterance. The NDL approach considers phonetic segments as the building block of L1 phonological and phonetic grammar. Applications of this

approach have achieved moderate success in predicting accentedness. However, questions remain as to whether it can predict accentedness of shorter utterances.

Experiments 1 and 2 have shown that stimuli with consonant and syllable mismatches were rated as more accented than stimuli with vowel mismatches. One possible explanation for such finding is that consonants are contextually less variable than vowels. Consonant mismatches, especially the ones that do not occur in L1 English, could therefore be more salient to listeners and consequently be judged as being more accented. The following section discusses Experiment 3, which adopted an NDL approach in its estimation of L1 and L2 differences. The experiment investigates whether raters' L1 knowledge affected their accentedness judgments.

6.4 The Experiment

Experiment 3 employed an NDL method to test whether L1 knowledge affects accentedness perception. IPA transcriptions from 100 L1 American English speakers were selected from the SAA to computationally approximate raters' L1 knoweldge. IPA transcriptions of the 100 L2 stimuli used in the two perception experiments were evaluated by the model to estimate the degree of dissimilarity between the L2 stimuli and their corresponding L1 productions. The degree of dissimilarity was termed the "NDL-distance." Accentedness ratings collected in Experiment 2 were compared against the NDL-distances to evaluate whether L1 knowlege could have affected accentedness judgment.

6.4.1 Materials

Experiments 1 and 2 focus on L2 productions in five phonological contexts namely "*please call*," "*ask her*," "*six spoons*," "*five thick*" and "*small plastic*." To investigate L1 productions of the five contexts, IPA transcriptions for 100 L1American English speakers were selected from SAA. For each L1 American English speaker, IPA transcriptions for the aforementioned five contexts were chosen, yielding 500 IPA transcriptions for L1 American English speakers (See Section A.2 of Appendix A for Demographic information of these 100 native speakers). Speech samples from L1 speakers of non-American L1 English varieties were not selected mainly because most of these speech samples were not yet transcribed.

6.4.2 Procedure

The 500 IPA transcriptions from the SAA were used to build a model that could potentially approximate L1 knowledge of the raters. Following Wieling et al. (2014b), every word was considered a lexical outcome. For each word, trigram combinations of segments were considered cues. For example, for the word "*ask*" and its associated pronunciation [æsk], the outcome is the word "*ask*," and its cues were represented as "*#æs*," "*æsk*" and "*sk#*," where the "#" symbol represents word boundaries. The trigram cues were constructed following recommendations by Baayen et al. (2016).

According to Baayen et al. (2016), there are three reasons for such a treatment. First, human perception of a phonetic segment is affected by its adjacent segments. Perception of voiceless plosives, for example, are discriminated by formant transitions of adjacent vowels in addition to acoustic properties of the plosives. In other words, the perception of voiceless plosives depends on both the stops and their adjacent segments, rather than the plosives alone. Using trigrams as cues is therefore potentially reflective of human perception of continuous speech.

Second, sequences of three segments reflect which segment combinations are permissible in English (i.e., English phonotactics). The analysis of trigrams, therefore, provide insights into phonotactic constraints of English. Third, learning algorithms using trigrams as cues have a lower chance of overfitting.[1] According to Baayen et al. (2016), learning algorithms such as the NDL have a higher chance of overfitting the data if unigrams (e.g., one segment) are used as cues. Baayen et al. (2016) recommended using the trigram structure as listed in the previous paragraph for the analysis of English, because it "*provides excellent discrimination without overfitting*."

Experiment 3 investigates sub-phonemic information by including diacritic marks in the analysis. A diacritic was not considered an independent segment, but part of a segment. For example, [æ] and [a̟] were treated as two independent segments. The inclusion of diacritic symbols in the model is based on previous findings that sub-phonemic information could potentially affect accentedness judgment. Trigram cues were constructed to account for English phonotactics. The NDL model also incorporated lexical information by calculating the association strength from trigram phonetic

[1] Overfitting is a modeling error which occurs when a function is too closely fit to a limited set of data points. When a model is too specific about a given dataset, it often fails to predict future observations.

cues to their corresponding lexical outcomes. Therefore, the NDL model adopted in Experiment 3 considers the phonetic, phonological and lexical information of the speech samples.

The whole model-building process was performed 100 times. Each time, IPA transcriptions from 50 L1 speakers were randomly chosen for building an L1 production model. It is not desirable to include speech samples from all the 100 L1 speakers in the model, because such treatment would increase the chance of overfitting. The L1 production model was constructed based on the associated strength from each cue (e.g. "*#æs*") to its lexical outcome (e.g. "*ask*"). The association strengths could be intuitively defined as the probability for a cue to associate with a certain outcome. Association strengths were calculated in R with the *estimateWeight* function in the *ndl* package (Arppe et al., 2018), using the Danks' equilibrium equation (Danks, 2003). Table 6.2 illustrates the association strength from each cue to its outcome.

Table 6.2: Association Strengths

Cues	Outcomes	Association Strengths
#æs	*ask*	0.166
æsk	*ask*	0.167
sk#	*ask*	0.667
#e̞s	*ask*	0.147
e̞sk	*ask*	0.147
#ɚ#	*ask*	1.000
#hɚ	*ask*	0.500
hɚ#	*ask*	0.500

Table 6.2 shows that "*sk#*" has a higher association strength than "*#æs*" or "*æsk*." It is understandable given the fact that the selected 100 L1 American English speakers did not always pronounce the "*a*" in "*ask*" as [æ]. They did, however, almost always pronounce the "*sk*" in "*ask*" as /sk/. In other words, the pronunciations for the "*a*" in "*ask*" were more variable than pronunciations of "*sk*." For the word "*her*," the consonant /h/ could be dropped in context "*ask her*". Therefore, the association strength from "*#ɚ#*" to "her" is 1.000, showing that "*#ɚ#*" has a 100% chance of predicting "her" in the context of "*ask her*". Given the calculated associated strengths, the association

strength from [æsk.ɚ] to the outcome "*ask her*" was calculated by summing association strengths of all the trigram cues (i.e., "*#æs*," "*æsk*," "*sk#*," "*#ɚ#*"") and then divide it by the number of words, which was 2 in this case. The association strength for [æsk.ɚ] is therefore:

$$(0.166 + 0.167 + 0.667 + 1.000) \div 2 = 1.000$$

For L2 productions containing cues that were not observed in L1 speech data, the association strengths for the unobserved cues were defined as 0. For example, L2 production [ask.hɚ] contains cues "*#as*," "*ask*," "*sk#*," "*#hɚ*," "*hɚ#*". Since "*#as*" and "*ask*" were not observed in L1 speech data, their association strengths were considered 0. The association strength from [ask.hɚ] to its outcome "*ask her*" is therefore:

$$(0 + 0 + 0.667 + 0.500 + 0.500) \div 2 = 0.834$$

The association strength for an L2 production could be intuitively interpreted as how much the L2 production meets the L1 listeners' expectations or how probable the L2 production could correctly convey its intended meaning to L1 listeners. For example, [ask.hɚ] has an 83.4% chance of conveying its intended meaning "*ask her*" to L1 listeners. This study assumed that the L1 pronunciations of "*ask her*" has an 100% chance of conveying its intended meaning to L1 listeners. In other words, the association strengths from L1 pronunciations to their corresponding lexical outcomes were defined as 1.000. The degree of dissimilarity between L2 production [ask.hɚ] and its corresponding L1 productions is therefore 0.166 (i.e., 1.000 – 0.834).

In some previous studies, similar dissimilarity measurements were often term phonetic distance or pronunciation distance (e.g., Baayen et al., 2016; Wieling et al., 2014b). To avoid confusion and to emphasize the fact that the NDL-based dissimilarity measurement was more than just phonetic dissimilarity, the current study opted to term the NDL-based dissimilarity measurement as "NDL-distance."

The NDL model was run on the data from 50 of the 100 L1 American English speakers. The model was run 100 times. Each time, a different set of 50 L1 American English speakers were randomly chosen to build a slightly different L1 production model, which generated a slightly different association strength for each trigram cue. Consequently, the NDL-distance between an L2 production and its corresponding L1 productions was slightly different each time the model estimation was

run. The averaged NDL-distances calculated by the NDL method across 100 runs were recorded for further analysis as approximations for dissimilarity.[2]

6.4.3 Results

The NDL-distance approximated the degree of dissimilarity between an L2 stimulus and the corresponding 100 L1 productions. The larger the NDL-distance, the more dissimilar an L2 stimulus is from L1 productions. The current study expects that the NDL-distance, as a measurement for the degree of dissimilarity between L1 and L2 speech, could predict the degree of foreign accentedness. Results from Experiment 2 show that stimuli with consonant and syllable mismatches were more accented than stimuli with vowel mismatches. However, if accentedness is purely decided by NDL-distances, then there should be no difference between the three types of mismatches once NDL-distances are controlled for. Alternatively, if consonant and syllable mismatches are perceptually more accented because of some inherent cognitive biases, then consonant and syllable mismatches will be more accented than vowel mismatches, even when NDL-distances are controlled for.

The following section first discusses the correlation between the NDL-distances and the acceted-ness ratings obtained by Experiment 2. It then discusses the linear mixed-effects models that further investigated whether stimuli with consonant and syllable mismatches are more accented than stimuli with vowel mismatches when NDL-distances are controlled for.

Correlation between Acccentedness Ratings and NDL-Distances

The NDL model implemented in Experiment 3 assumes that lexical outcomes of the trigram cues are known to the listeners. Data from Experiment 2 were therefore chosen for the evaluation of the NDL-distances, because raters of Experiment 2 had a clear expectation of the lexical outcome for each stimulus before the stimulus was heard. Figure 6.1 below illustrates the mean accentedness ratings for the 100 stimuli obtained in Experiment 2 against the NDL-distances. The gray dots represent accentedness ratings against NDL-distances. The solid line represents the fitted linear regression line. Mean accentedness ratings represent the accentedness judgments on the stimuli. On the y-axis,

[2]For the illustration of this NDL model, a web application was made by the current study and it is available on `https://gaozhiyan.shinyapps.io/ndl_calculator`

the larger the accentedness rating, the more accented a stimulus was judged. The NDL-distances on the x-axis approximate the degree of dissimilarity between an L2 stimulus and its corresponding L1 productions.

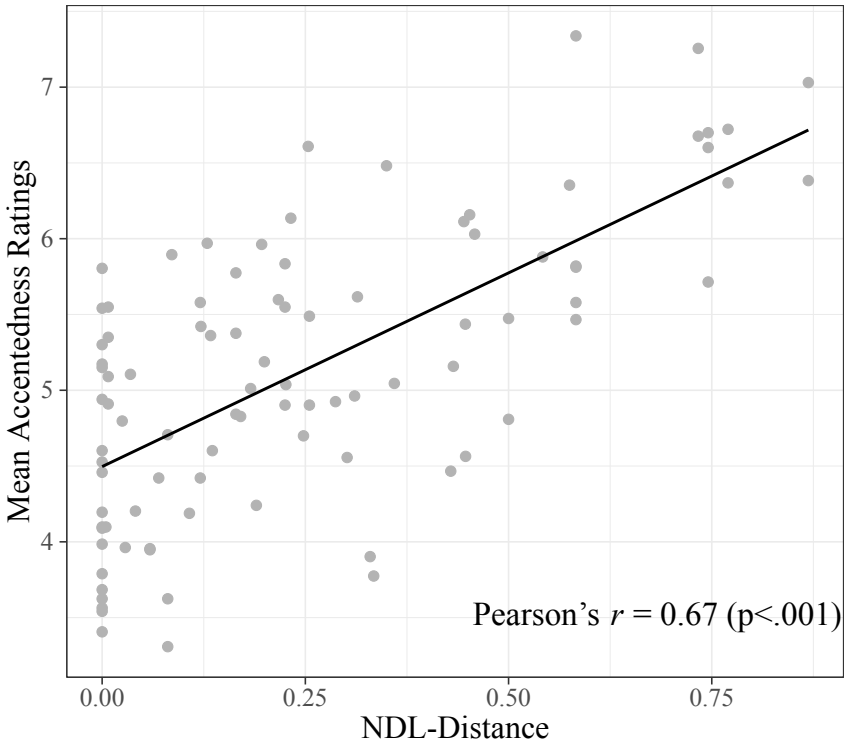

Figure 6.1: Relationship between Accentedness and NDL-Distance

Overall, there is a clear positive relationship between the accentedness ratings and the NDL-distances (Pearson's r = .67, p <.001). This result generally agrees with the prediction that the larger the NDL-distance between an L2 stimulus and its corresponding L1 productions, the more accented the L2 stimulus will be judged.

Types of Mismatches

Linear mixed-effects regression models were constructed to investigate the effects of the four types of stimuli (i.e., consonant vs. syllable vs. vowel vs. match), while controlling for NDL-distances.

Rating data from Experiment 2 were used for the analysis. Accentednss ratings were the dependent variable. Helmert-contrast-coding was again performed on the "types of stimuli" variable to create three contrasts. The first contrast compares stimuli with consonant mismatches and stimuli with syllable mismatches, the second contrast compares vowel mismatches with consonant and syllable mismatches, and the third contrast compares the match stimuli with the three types of mismatch stimuli. The three contrasts, trial numbers and NDL-distances for the stimuli were entered as fixed effects. Two-way and three-way interactions of the fixed effects were also entered as fixed effects. Raters and stimuli were entered as random effects.

Model comparisons using likelihood ratio tests showed that (1) NDL-distances significantly contributed to model fit (β = 1.74, $\chi 2$ = 8.79, p<.01), showing that the degree of dissimilarity between L1 and L2 speech samples could have affected accentedness perception; (2) trial number contributed significantly to model fit (β = 0.6, $\chi 2$ = 72.24, p<.001), showing that the accentedness ratings increased overtime; and (3) the three contrasts did not contribute significantly to model fit, showing that accentedness differences between the four types of stimuli diminished once NDL-distances were controlled for.

To further investigate the interactions between NDL-distances and the types of mismatches, mean accentedness ratings of each type of mismatch were plotted against NDL-distances. Figure 6.2 demonstrates the relationship between NDL-distances and accentedness ratings of stimuli with consonant and syllable mismatches.

Figure 6.2 shows that the mean accentedness ratings of different types of consonant mismatches positively correlates with NDL-distances (Pearson's r = 0.58, p<.01). For example, the two instances of VOT-shortening on consonant cluster /pl/ in the context of "*please call*" had an NDL-distance as large as 0.886 and the two instances were judged as being relatively more accented. The instance for VOT-shortening on /pl/ in the context of "*small plastic*" had an NDL-distance as small as 0.070, and it was judged as being less accented.

There are, however, a few discrepancies. The NDL-distance failed to model accentedness for stimuli with non-English phonemes. For example, one of the four cases for /ɹ/-trilling (i.e. /ɹ/ →/r/)

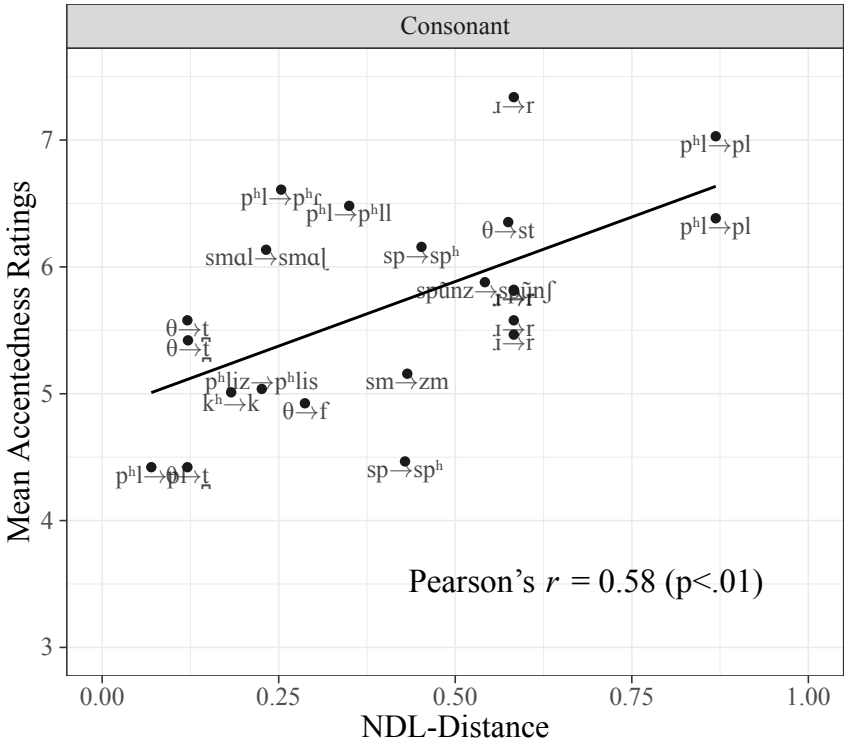

Figure 6.2: Relationship between Accentedness and NDL-Distance (Consonant)

was assigned a relatively higher accentedness rating, but its NDL-distance was not as large as expected. The use of the retroflex [ɭ] such as [smɑl]→[smɑɭ] was also rated as very accented. Possible reasons for the discrepancies are discussed in the Discussion section of this chapter. However, NDL-distances for these stimuli were smaller than expected. NDL-distances did correlate well with accentedness ratings of stimuli with consonant mismatches.

Figure 6.3 shows the relationship between NDL-distances and ratings of stimuli with syllable mismatches. There is also a clear positive correlation between NDL-distances and accentedness ratings of stimuli with syllable mismatches (Pearson's r = 0.69, p<.001). For example, the NDL-distances for stimuli with vowel epenthesis are around 0.750, while the NDL-distances for stimuli with segment deletion are generally lower than 0.500. Accentedness ratings show that vowel

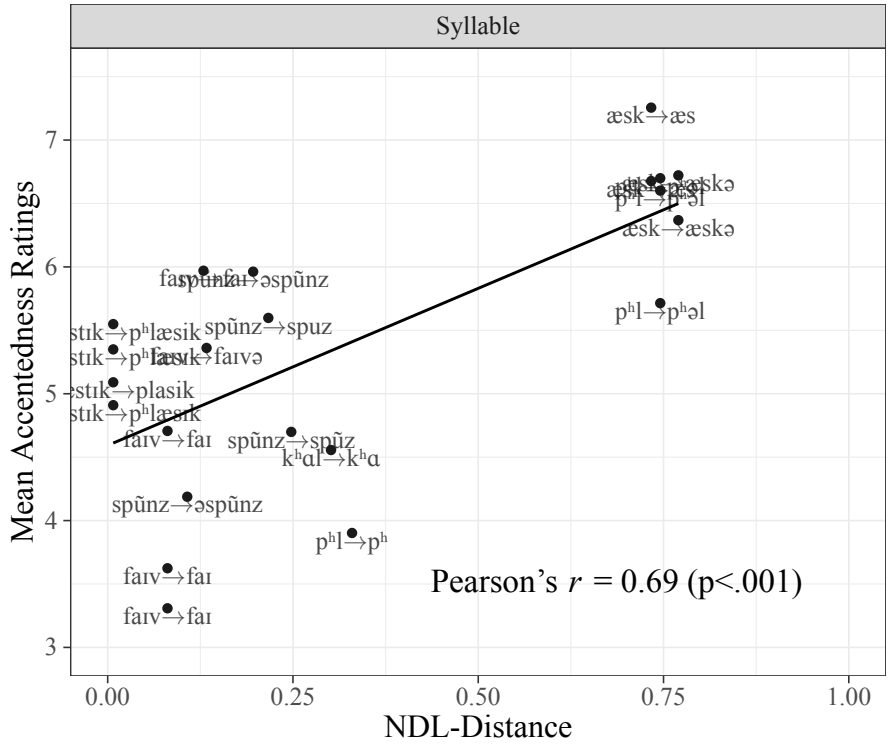

Figure 6.3: Relationship between Accentedness and NDL-Distance (Syllable)

epenthesis was indeed judged as being more accented than segment deletion. Interestingly, /k/-deletion in the context of "*ask her*" was rated as relatively more accented than other types of consonant deletion. This particular exception is also captured by the NDL-distance measurement.

The correlation between NDL-distances and accentedness ratings of vowel mismatches is also positive (Pearson's r = 0.44, p<.05). Figure 6.4 illustrates the relationship between NDL-distances and mean accentedness ratings of stimuli with vowel mismatches.

Dialectal variations such as [ɑ]→[ɔ] in words "*small*" and "call" were assigned smaller NDL-distances. Non-dialectal variations such [aɪ]→[a] and [aɪ]→[ɑɪ] in word "*five*" were assigned relatively larger NDL-distances. Dialectal variations indeed received relatively lower ratings than did non-dialectal variations. Therefore, the NDL-distance explains rating differences between vowel

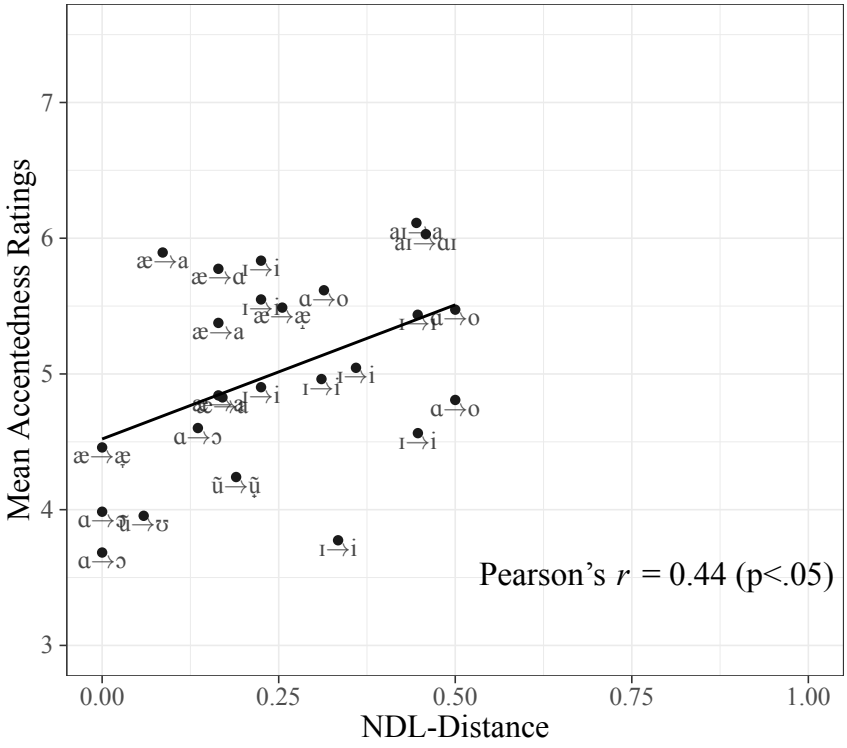

Figure 6.4: Relationship between Accentedness and NDL-Distance (Vowel)

mismatches to a certain degree. However, the range of NDL-distances for stimuli with vowel mismatches is restricted to the range from 0 to 0.500. It is known in the field of statistics that a restricted range severely affects correlation. Therefore, the Pearson correlation calculated above might not be meaningful. The reason for the restriction of NDL-distances (i.e., a narrower range) was because vowels of the L1 productions exhibit a relatively higher degree of variation, especially when diacritics are considered. The NDL algorithm therefore assigned lower association strength to trigram cues involving vowels. In other words, vowels in an utterance did not contribute as much to the association strength as did consonants, unless the vowels are epenthetic, which was considered an issue of syllable mismatch. As a result, the range of NDL-distances for vowel mismatches is limited to under 0.500. In order to fully examine the possible effect of NDL-distances on accentedness of vowel mismatches, a wider range of NDL-distance is required.

6.4.4 Summary

Results of Experiments 1 and 2 show that the frequency of occurrence of a speech pattern in L1 speech could potentially affect the accentedness judgment on the phonetic pattern. For example, pronouncing "*thick*" as [tɪk] or [fɪk] was observed in L1 speech data. L2 stimuli involving pronouncing "*thick*" as [tɪk] or [fik] were judged as less accented. On the other hand, pronouncing "*ask*" as [æsk] was not observed in L1 speech data. [æskə] was indeed judged as relatively more accented. These results indicate that the raters were aware of which speech patterns are allowed in L1 speech and have made their accentedness judgment based on such knowledge. Experiments 1 and 2 made a general claim that consonant mismatches are in general more accented than syllable or vowel mismatches. However, frequency of occurrence of a speech pattern in L1 speech could have potentially skewed the result. Experiment 3 described in the current chapter implements an NDL approach to account for the raters' phonetic and phonological knowledge with regard to the five contexts.

Experiment 3 constructed an L1 production model and subsequently measured the NDL-distance between L2 productions and their corresponding L1 productions. The calculation was based on the notion of association strength, which was defined as the probability that a trigram phonetic cue (e.g., *#æs*, *æsk*, *sk#*) is associated with a certain lexical outcome (e.g., "*ask*"). The results show that the accentedness differences between consonant, syllable and vowel mismatches diminished when NDL-distance was controlled for. In other words, rating differences between the various types of mismatches can be explained by how much they differ from their respective L1 productions.

Results in this chapter show that the larger the NDL-distance between an L2 production and its corresponding L1 productions, the more accented the L2 production is perceived. This trend was particularly evident when consonant and syllable mismatches were concerned. The effect of NDL-distance on stimuli with vowel mismatches was not as clear. A possible reason for this discrepancy can be attributed to the relatively weaker association strength of a vowel in an utterance to its lexical outcome. The results show that the vowel mismatches were not necessarily judged as less accented than consonant or syllable mismatches once NDL-distance was controlled for. It is therefore possible that the L1 knowledge, as approximated by the NDL model, could be responsible for accentedness

judgment.

6.5 Discussion

Foreign accent, as Major (2012) defined it, is a pronunciation deviating from what an L1 speaker expects another L1 speaker to sound like. The current study therefore first built an L1 production model to estimate what an L1 speaker should sound like. Association strengths were used to calculate the degree of difference between L1 and L2 speech samples. The NDL approach implemented by Experiment 3 is advantageous in several aspects. First, it is based on Rescorla and Wagner (1972)'s learning theory, which potentially reflects human cognition, making it more cognitively grounded than other alignment methods. Second, the NDL approach built an L1 production model based on productions of 100 L1 speakers of American English, rather than codified pronunciations in a dictionary. The L1 production model therefore potentially reflected sound changes in certain phonological contexts and dialectal variations of American English. Third, the NDL approach incorporated lexical outcomes into its algorithm, which reflects the experimental design of Experiment 2. The NDL-distance does, to some degree, explain empirical perception data. One could therefore tentatively draw the conclusion that perceptual foreign accentedness is related to association strengths from trigram cues to lexical outcomes.

The current study investigated the effect of NDL-distance on different types of stimuli. Results show that the effect of NDL-distance is more evident on stimuli with consonant and syllable mismatches than on stimuli with vowel mismatches. Analysis of consonant mismatches revealed a few discrepancies. For example, stimuli with non-English phonemes such as the retroflex [ɭ] and the trill [ɹ], were judged as being very accented. Their corresponding NDL-distances are not as large as expected. A potential reason for such a discrepancy lies in the variability of consonants and vowels in L1 speech. For pronunciations of the word "*small*," The NDL model apparently assigned lower association strength to trigram cues involving the /l/ in "*small*" (i.e., "*mal*" and "*al#*"). L1 productions for the vowel in the word "*small*" could be [ɑ, ɔ, a, ɑ̟, ɑ̠, ɑː, ɔː, aʊ, ɑʊ̆] according to IPA transcriptions in the SAA. On the other hand, the "sm" in "*small*" was always produced as [sm]. Consequently, trigram cue "*#sm*" was assigned a higher association strength than "*sma*" or "*al#*," simply because

"*#sm*" always associated with "*small*," while "*sma*" and "*al#*" did not. In other words, replacing the English /l/ in "*small*" with a retroflex [ɭ] has only a limited effect on association strength. However, raters of the two perception studies judged the retroflex [ɭ] as being very accented. In order to better approximate accentedness judgment, an improved algorithm should be able to reduce association strengths to a larger degree once a non-English sound is found.

The NDL-distances of stimuli with vowel mismatches were lower than 0.50, which might have resulted from the fact that vowels tend to be more variable than consonants in L1 speech. As mentioned previously, L1 productions of the vowel in word "*small*" could be [ɑ, ɔ, a, ɑ̟, ɑ̠, ɑː, ɔː, aʊ, ɑŏ], according to IPA transcriptions in the SAA, while the consonants in the word "*small*" were almost always produced as [s], [m] and [l]. The source for the variability of vowels could be either dialectal or phonological assimilation. Consonants in L1 speech also vary depending on dialects and phonological contexts. However, consonants seem to be less variable than vowels, as far as the data of the current study are concerned.

Experiment 3 did not utilize acoustic information of L1 and L2 speech in the approximation of NDL-distance. The reason for the omission of acoustic comparison is that the L2 stimuli selected by the current study are limited in their types and tokens. Acoustic information of phonemes is multidimensional. For example, potential acoustic correlates of English plosives include but are not limited to burst intensity, closure and VOT durations. Various measurements for phonation might also be relevant. According to Lisker (1986), 16 different phonetic cues could be relevant to the distinction between /pa/ and /ba/. Although it is possible to measure all these acoustic properties, questions remain as to which ones are perceptually relevant. To answer these questions, stimuli should be more carefully designed so that acoustic correlates of phonemes could be uncovered. The 100 stimuli selected by the current study are not suitable for such a task.

A potential problem of the NDL method in Experiment 3 is that the "association strength" of trigram cues was calculated based on L1 speakers' productions. L2 speech might involve trigram cues that do not exist in L1 speech. As mentioned above, some L2 English speakers pronounced the coda /l/ in "*small*" as a retroflex [ɭ], which was not observed in L1 speech data. The NDL algorithm provided limited insight into the association strength from the L2 utterance [smɑɭ] to its lexical

outcome "*small*". Experiment 3 assigned 0 to the association strengths of trigram cues involving the retroflex [ɭ]. Results of such a treatment, however, did not correlate well with accentedness ratings.

An alternative approach of modeling could consider distinctive phonetic features of segments. The Maximum Entropy approach (MaxEnt: Hayes and Wilson, 2008), for example, utilizes distinctive phonetic features to model L1 knowledge. The retroflex /ɭ/, although does not exist in English, shares all its phonetic features with English phonemes. By measuring the co-occurrence of phonetic features in English, the phonetic distance from /ɭ/ to its L1 target /l/ could be estimated. The potential problem with the MaxEnt method is that it assumes full feature specification of phonemes, which might not be reflective of speech perception. Instead of full feature specification, the ALINE method (Kondrak, 2003) considers only major natural classes (e.g., place, syllabic, nasal, high, back, etc.). The natural classes can be weighted using gradual learning algorithms such as the MaxEnt model. However, empirical studies are still needed to investigate whether and how major natural classes affect perception

In addition to not being able to account for non-English phonemes, the NDL method in Experiment 3 seems to have a problem of assuming a direct mapping from speech sounds to lexical items. Chomsky's Minimalist program claims that phonological/phonetic processing at the phonetic form (PF) and semantic processing at the logical form (LF) do not directly exchange information, but are mediated by syntax (Chomsky, 1995). There is no doubt that the direct mapping from speech sounds to lexical items has oversimplified language processing. Syntax and morphorsyntax need to be taken into consideration. For example, among the five phonological contexts selected by the current study, the contexts "*five thick*" and "*small plastic*" are not necessarily grammatical. How the degree of grammaticality interacts with accentedness perception was ignored by the current study. Although the NDL method did not consider syntax or morphorsyntax, it mapped pronunciation to lexical items through occurrences of segment sequences (i.e., the trigram cues). The occurrences of segment sequences potentially reflect rules of English phonotactics, which is part of L1 phonological grammar. The current study wishes to claim that the association strengths, as calculated by the NDL method, are a part of L1 knowledge that affects accentedness judgment, while fully acknowledge that the measurement of association strength is but a simplified representation of language processing.

Chapter 7: Conclusion

While much research has investigated the accentedness of non-native (L2) English speech, very few has attempted to rank individual phonetic and phonological patterns in L2 speech according to their perceived foreign accentedness. The current study contributes to the field of foreign accent by providing accentedness rankings of various phonetic patterns in L2 speech. And even while the previous studies in foreign accent that did focuss on specific phonetic patterns in L2 speech only dealt with a limited few, the current study investigated the perceived accentedness of a larger variety of phonetic patterns.

Based on empirical observations in Experiments 1 and 2, the current study further hypothesized that raters' knowledge of English phonetics and phonological has affected their acccentedness judgment. A Naïve Discriminative Learning (NDL) model was built to evaluate this hypothesis (Chapter 6). The NDL model achieved moderate success in explaining rating differences between different types of L2 stimuli. This chapter will first summarize major findings of the current study and discuss their implications. The focus will then shift to limitations of the current study and recommendations for future research.

7.1 Summary of Results

The current study selected 100 L2 English speech samples from the Speech Accent Archive (SAA) for two perception experiments to elicit accentedness judgment on different types of phonetic patterns. A third experiment was conducted to investigate how raters' L1 knowledge affects their accentedness judgement. In Experiments 1 and 2, native (L1) American English listeners (i.e., the raters) listened to and rated the foreign-accentedness of 100 L2 speech samples. Experiment 1 and 2 focused on the observation of accentedness ratings of the L2 speech samples in five phonological contexts, namely, "*please call*," "*ask her*," "*five thick*," "*six spoons*" and "*small plastic*." Phonetic

transcriptions (IPA transcriptions) from 100 L1 American English speakers in the SAA were surveyed to find the most common L1 productions for the five contexts (e.g., [faɪv θɪk] for "*ask her*").

L2 speech samples whose IPA transcriptions match the most common L1 productions were termed match stimuli. L2 speech samples whose IPA transcriptions that differ from the most common L1 productions were termed the mismatch stimuli. Among the mismatch stimuli, some differ from the most common L1 productions by only one consonant (e.g., [faɪv tɪk]), some differ from the most common L1 productions by only one vowel (e.g., [faɪv θik]), some differ from the most common L1 productions by only syllable structural element (e.g., [faɪvə θɪk]). These three types of stimuli were termed consonant mismatch, vowel mismatch, and syllable mismatch, respectively. Some of the mismatch stimuli contain phonetic patterns that were not observed in L1 speech (e.g., [faɪvə θɪk], [faɪv θik]); others contain phonetic patterns that exist in L1 speech (e.g., [faɪv tɪk]). Acoustic analysis was conducted to verify the reliability of the IPA transcriptions for the L2 speech samples. The current study accepted the IPA transcriptions as reliable because acoustic differences between the L2 speech samples and their corresponding L1 speech samples were captured by the IPA transcriptions for the L2 speech samples.

The results of the two perception experiments show that stimuli with consonant and syllable mismatches were judged as being more accented than stimuli with vowel mismatches. The three types of mismatch stimuli were judged as being more accented than the match stimuli. The types of mismatches that exist in L1 speech were judged as relatively less accented (e.g., [θ]→[t]). The types of mismatches that do not exist in L1 speech were judged as relatively more accented (e.g., [θ]→[s̪t̪]). These results show that raters were aware of which phonetic patterns are allowed in L1 speech. Analysis also show that phonological contexts could have affected the accentedness judgment on some types of mismatches. For example, VOT-shortening was rated as being more accented phrase-initially than phrase-medially. The two perception experiments therefore show that (1) different types of phonetic patterns in L2 speech do not carry equal weight in accentedness perception; (2) raters' knowledge of L1 phonetics and phonology potentially affects accentedness judgment; and (3) phonological context potentially affects accentedness perception.

The two perception experiments differed in their respective experimental designs. Experiment 1

did not provide any training to the raters. In addition, the raters in Experiment 1 were not aware of the intended meanings of each stimulus. Unlike Experiment 1, Experiment 2 included a training phase, and it controlled for intelligibility by informing raters of the intended meaning of each stimulus. As a result, ratings obtained in Experiment 2 were more consistent than ratings obtained in Experiment 1. The stimuli were generally judged as more accented in Experiment 2 than in Experiment 1. Accentedness ratings of both Experiment 1 and Experiment 2 experienced a gradual increase during the entirety of the experiment. That is, the same stimulus might receive a higher rating (i.e., be perceived as being more accented) if it occurred later in the experiment. These results show that (1) a training phase might be valuable to familiarize raters with the procedure of the experiment and the range of the accents covered by the stimuli; (2) stimuli tend to be judged as being more accented once their intended meanings were known.

Results of the two perception studies indicate that raters' L1 phonetic and phonological knowledge with regard to the five phonological contexts (L1 knowledge) could have affected their accentedness judgment. Experiment 3 attempted to directly investigate how raters' judgments of accentedness were affected by their L1 knowledge. A Naïve Discriminative Learning Model (NDL) was employed to investigate raters' L1 knowledge by examining the co-occurrences of pronunciations (e.g., [æsk]) and lexical outcomes (e.g., "*ask*") in the L1 speech of American English. The model assigned association strengths to trigram sequences (e.g., "*#æs*," "*æsk*," "*sk#*"), measuring the probability for each trigram to predict the intended lexical outcome (e.g., how probable it is for "*#æs*" to predict "*ask*"). Raters' L1 knowledge was, therefore, approximated by using a matrix of association strengths, mapping trigram phonetic segment sequences to lexical outcomes.

The 100 L2 stimuli were subsequently evaluated against raters' L1 knowledge to estimate the degree of dissimilarity (NDL-distance) between L1 and L2 speech samples. The results showed that NDL-distance correlates significantly with accentedness ratings obtained by Experiment 2, suggesting that L1 knowledge, as approximated by the NDL model, could have potentially affected accentedness judgments. Comparisons between linear mixed-effects regression models further revealed that rating differences between stimuli with consonant, vowel and syllable mismatches were not

significant when NDL-distances were controlled for. These results show that rating differences between the different types of stimuli could be explained by the association strengths from the stimuli to their intended lexical outcomes.

The NDL analysis reveals that consonants in the 100 L1 American English speakers' speech do not vary as much as vowels. For example, L1 productions of the vowel in such a word as "*small*" could be [ɑ, ɔ, a, ɑ̟ , ɑ̠, ɑː, ɔː, aʊ, ɑʊ̈], according to IPA transcriptions in the SAA. By comparision, the consonants in such a word as "*small*" exhibit a considerably lower degree of variability. The NDL model, therefore, assigned higher association strengths to trigram cues involving consonants than to trigram cues involving vowels. Changing the consonants in a pronunciation is therefore more likely to lower its association strength to its lexical outcome. In other words, consonant changes are more likely to impair lexical identification. Previous studies in lexical identification claim that consonants are more important than vowels in identifying lexical outcomes (Nespor, Peña, and Mehler, 2003). Findings of Experiment 3 support such a claim and further demonstrates that lexical identification potentially plays a role in accentedness judgment.

7.2 Theoretical Implications and Societal Impacts

The current study has important theoretical and applied implications. L1 English listeners' perception of foreign accent reveals the nature of "foreign accentedness" and how the deviation from L1 grammar affects accentedness perception. Sociolinguistics research often reports that pronunciation, rather than vocabulary or syntax, is a major factor that affects communication (Grant and Brinton, 2014). Although L2 accents bear no relationship to one's intelligence or personal character, they do carry a potential stigma that could cause negative social and workplace consequences (Gluszek and Dovidio, 2010). Second language or foreign language learners, therefore, place great importance on the correctness of their pronunciation (Waniek-Klimczak, Rojczyk, and Porzuczek, 2015). However, many English language instructors are reluctant to incorporate pronunciation instruction into their teaching curriculum (Thomson, 2014). One reason for such reluctance is that L2 pronunciation errors are numerous, and there is not enough time for teachers to address all of them (Munro and Derwing, 2006; Thomson, 2014).

By identifying phonetic patterns that are most accented to L1 English listeners, the current study could help language teachers set priorities for pronunciation instructions. The current investigation therefore could enable a much more efficient and perhaps effective strategy for pronunciation instructions and could more broadly help disadvantaged linguistic minorities to achieve their full potentials in society. The current research also endeavored to model L1 productions computationally and subsequently compare L2 productions against L1 productions. The model adopted by the current research potentially reveals the nature of "foreign accentedness" and could further help design improved speech analysis algorithms.

7.3 Discussion and Future Directions

The current study based its analysis on IPA transcriptions rather than acoustic information of the selected speech samples. Therefore, the reliability of the IPA transcriptions was of utmost importance. Although transcribers of the SAA, from which data of the current study were extracted, were diligent in transcribing every speech sample, inaccuracies are still possible. The current study, therefore, examined the IPA transcriptions for the 100 selected stimuli by measuring benchmark acoustic signals. Results of the acoustic analysis have indeed partially validated the transcriptions, but arguments could still be made against the validity of the acoustic analysis itself. As stated in Chapter 3, acoustic correlates of phonemes are multidimensional. The current study examined only one or two acoustic measurements of relevant segments, and this could have oversimplified the issue.

On the other hand, acoustic analysis mostly concerns speech production, while IPA transcriptions in the SAA could be considered as a reflection of the transcribers' perception of the speech samples. To increase the reliability of the IPA transcriptions, the most direct approach is perhaps to recruit more trained transcribers. When there are more people transcribing the same speech samples, the most reliable transcriptions could then be determined via inter-transcriber agreement. We have undertaken such an endeavor. Preliminary results showed that 73% of the transcriptions submitted by the 67 newly recruited transcribers matched the existing transcriptions in the Speech Accent Archive (Weinberger et al., 2019). More trained transcribers are still needed to further improve the reliability of the IPA transcriptions.

Previous research in foreign accent often suggests that the degree of (dis)similarity between an L2 speech and its L1 counterparts is responsible for the degree of perceived foreign accentedness. The practical problem in comparing L1 and L2 speech samples is that L1 speech exhibits considerable within- and between-speaker variability. The term "native speaker norm" is often mentioned to refer to speech patterns of a hypothetical "average" L1 speaker. The current study adopted a similar concept in its selection of stimuli by classifying stimuli based on whether they match the most common L1 pronunciations in the SAA.

Pronunciations that differ from the most common L1 productions were considered as containing mismatches. The potential drawback is that some of the mismatches could be dialectal variations of L1 English (e.g., pronouncing "*thick*" as [tɪk] or [fɪk]). These variations indeed received relatively lower accentedness ratings (i.e., less foreign accented). The general finding in Experiments 1 and 2 is that consonant mismatches are more accented than vowel mismatches. However, if more L1 dialectal vowel variations were included in the current study than L1 dialectal consonant variations, then the general finding should be that non-dialectal variations are more accented than dialectal variations. To investigate whether consonants are indeed more important than vowels in accentedness judgment, we need to consider how likely for a specific mismatch to exist in L1 speech.

Experiment 3 further investigated how the frequency of occurrence of a phonetic/phonological pattern in L1 speech could affect the accentedness judgments on this phonetic/phonological pattern. The model assigned higher association strength to phonetic patterns that could be considered dialectal (i.e., spoken by a subset of the 100 L1 speakers). For example, only two out of the 100 L1 speakers pronounced "*thick*" as [tɪk]. The association strength for [tɪk], as calculated by the NDL model, is 73.27% rather than 2%, showing that [tɪk] has a 73.27% probability of predicting the lexical outcome "*thick*." Therefore, it was considered relatively more native-like. When the frequency of occurrence of a phonetic/phonological pattern in L1 speech was controlled for, the rating differences between consonant mismatches and vowel mismatches diminished. This finding demonstrates that consonant mismatches are not inherently more accented than vowel mismatches. Rating differences between the different types of stimuli could be explained via "association strengths" from phonetic segment sequences to their respective lexical outcomes.

Although the NDL model achieved moderate success in accounting for the data in the current study, the model was built on data from only 100 L1 speakers of American English, which are probably not representative enough. In addition to the omission of other L1 varieties of English, the 100 L1 speakers of American English were from 37 states and the District of Columbia. States such as Arizona, Colorado and Delaware were not represented (See Section A.2 in Appendix A for detailed demographic information). To better model raters' L1 knowledge, data from more L1 speakers are undoubtedly needed.

With more data from L1 English speakers, the model could be tailored to approximate accentedness judgments of specific groups of listeners. For example, L1 British English listeners' accentedness judgments could be approximated by including more production data from L1 English varieties in the British Isles. Some research shows that the experience with a certain type of accent could potentially affect accentedness judgment (Wester and Mayo, 2014). Mandarin listeners, for example, tend to judge Mandarin-accented English less accented than do L1 English listeners (Wester and Mayo, 2014). The NDL model could therefore potentially approximate Mandarin speakers' accentedness judgment by including production data from Mandarin speakers of English.

The current study did not discuss sociolinguistic elements such as one's own dialect and familiarity with L2 speech, both of which could potentially affect accentedness judgments. As shown in van den Doel (2006), British English speakers and American English speakers do not always agree on which phonetic patterns in L2 speech are accented. The current study focuses only on ratings from L1 listeners of American English. However, the L1 American English listeners' accentedness judgments are hardly homogeneous (See Appendix C for mean ratings from raters of each state).

L1 listeners of southern American English or people who are familiar with southern American English might be more tolerant to monophthongizations such as [aɪ] to [a], because such sound change is similar to off-glide deletions in many varieties of southern American English (Labov, Ash, and Boberg, 2005). Raters from regions with a large Hispanic population (e.g., California, Arizona and Texas) are very likely to have been exposed to Spanish-accented English, and thus could be more familiar with Spanish speakers' L2 English speech patterns such as vowel prothesis of s-clusters. Due to the limited access to raters' personal information, the current study could

not warrant a detailed investigation on sociolinguistic factors. Future research is needed to further investigate how one's sociolinguistic background affects his/her accentedness judgment.

Overall, findings in this dissertation contribute to the field of foreign accent by providing accentedness rankings of different types of phonetic/phonological patterns in L2 speech. This disseration further investigated why these phonetic/phonological patterns were weighted differently in L1 accentedness judgment. Results show that L1 knowledge, as measured by statistical properties of phonetic segments in L1 English speech, could have contributed to the degree of perceived accentedness.

Appendix A: Speaker Demographics

A.1 Non-native Speaker Demographics

The 100 audio stimuli were extracted from 93 different non-native English speakers. Information of the non-native speakers' gender, age, age onset of English learning (AO), duration of residence in an English-speaking country (LoR), way of English learning (style) and native language (L1) are available on the website of the Speech Accent Archive.

Among the 93 non-native English speakers, 41 of them are female and the remaining 52 are male. Age range is between 18 and 69. The mean age is 31.66 (SD=11.39). The AO ranges from 1 to 39. The mean AO is 13.60 (SD=7.87). The LoR ranges from 0 years to 41 years. The mean LoR is 6.28 (SD=8.30). 75 of the 93 non-native speakers learned English in academic settings. 18 of them reported to have acquired English in naturalistic ways.

The non-native speakers differed in their native languages. There are 53 different native languages represented. Table A.1 lists the 53 languages. The numbers in parentheses show the number of speakers of the listed languages.

Table A.1: L1 Background of the L2 English Speakers

[1]	afrikaans (1)	agni (1)	albanian (2)	amharic (1)
[5]	arabic (4)	armenian (1)	baga (1)	belarusan (1)
[9]	bengali (1)	bosnian (1)	bulgarian (1)	cantonese (5)
[13]	catalan (1)	czech (3)	danish (1)	dutch (1)
[17]	farsi (2)	french (4)	georgian (2)	german (4)
[21]	greek (1)	gujarati (1)	italian (3)	japanese (5)
[25]	kiswahili (2)	korean (2)	lamaholot (1)	mandarin (2)
[29]	mongolian (1)	norwegian (3)	polish (1)	portuguese (4)
[33]	pulaar (1)	punjabi (2)	quechua (1)	romanian (1)
[37]	russian (2)	satawalese (1)	serbian (1)	sinhalese (1)

[41]	spanish (4)	swedish (2)	swiss german (1)	taishan (1)
[45]	taiwanese (2)	tamil (1)	telugu (1)	thai (1)
[49]	tibetan (1)	turkish (1)	uzbek (1)	vietnamese (2)
[53]	zulu (1)			

A.2 Native Speaker Demographics

Chapter 6 selected 100 native speakers of American English for the construction of the native speaker model.

Among the 100 native speakers, 39 are female and 61 are male. Age range is between 18 to 79. The mean age is 36.65 (SD=16.93). 97 speaker reported that they reside in the United States. One speaker resides in the United States Virgin Islands. Two speakers reside in both the United States and the United Kingdom.

The 100 native speakers are from 37 states and the District of Columbia in the United States. Table A.2 lists the birthplaces of the native speakers. The numbers in parentheses show the number of speakers from the listed states.

Table A.2: Birthplace of the Native Speakers of American English

[1]	alabama (3)	alaska (1)	arkansas (3)
[4]	california (9)	connecticut (2)	district of columbia (1)
[7]	florida (4)	georgia (2)	idaho (1)
[10]	illinois (2)	indiana (3)	iowa (1)
[13]	kansas (1)	kentucky (1)	louisiana (3)
[16]	maine (1)	maryland (4)	massachusetts (3)
[19]	michigan (4)	minnesota (4)	mississippi (1)
[22]	missouri (1)	new hampshire (1)	new jersey (1)

[25]	new york (8)	north carolina (3)	ohio (7)
[28]	pennsylvania (5)	rhode island (1)	south carolina (1)
[31]	south dakota (1)	texas (4)	us virgin islands (1)
[34]	utah (1)	virginia (4)	washington (1)
[37]	west virginia (1)	wisconsin (3)	

Appendix B: Rater Demographics

B.1 Demographics Questionnaire

Please provide us with some information about you and how you did the experiment. We will keep this information private (it will not be associated with your worker id), and it will help us very much when we analyze the data.

1. Gender:

   ☐ male ☐ female

2. Age: _____

3. List your native language _____

   Note: Your native language is the language you heard, grew up with, and used from a young age, typically between the ages 0 and 4. It continues to be the language you identify with, and it is considered your first language.

4. List any other languages you speak _____

5. Please tell us your occupation _____

6. Please tell us your birth place (city/state/country) _____

7. Please tell us your current residence (city/state/country) _____

8. Have you ever had any speech or hearing disorders (e.g., Hearing loss, aphasia, dysphasia etc.)? ☐ yes ☐ no

9. If Yes, please explain.

   _____

10. Please provide us with any comments or suggestions you might have regarding this experiment.

    _____

B.2 Experiment 1: Rater Demographics

Experiment 1 recruited 110 raters. Responses from 108 of the 110 raters were used for the final analysis. As mentioned in Section 4.4 of Chapter 4, 61 of the raters are female, 45 are male, and two did not report their gender. The age of the 108 participants ranges from 20 to 66. The mean age is 33.50 (SD=12.51). Table B.1 lists the current residences and birthplaces of these 108 raters. The raters are from 33 states in the United States. The "Numbers (CR)" column lists the number of raters according to the raters' current residences. The "Numbers (BP)" column lists the number of raters according to the raters' birthplaces.

Table B.1: Raters' Current Residences and Birthplaces (Experiment 1)

Current Residences	Numbers (CR)	Birthplaces	Numbers (BP)
arizona	4	arizona	4
arkansas	1	arkansas	1
california	17	california	17
colorado	1	colorado	1
connecticut	1	connecticut	1
florida	8	florida	8
georgia	4	georgia	4
illinois	1	illinois	1
indiana	1	indiana	1
iowa	1	iowa	1
kansas	2	kansas	2
kentucky	2	kentucky	2
louisiana	1	louisiana	1
maine	1	maine	1
maryland	1	maryland	1
massachusetts	4	massachusetts	4

michigan	2	michigan	2
minnesota	1	minnesota	1
missouri	3	missouri	3
nevada	2	nevada	2
new hampshire	1	new hampshire	1
new jersey	6	new jersey	6
new mexico	1	new mexico	1
new york	11	new york	11
north carolina	5	north carolina	5
ohio	3	ohio	3
oklahoma	1	oklahoma	1
oregon	2	oregon	2
pennsylvania	4	pennsylvania	4
south carolina	2	south carolina	2
tennessee	2	tennessee	2
texas	7	texas	7
washington	5	washington	5

B.3 Experiment 2: Rater Demographics

155 raters were recruited for Experiment 2. Responses from 133 of the 155 raters were used for final analysis. Only the 133 raters' information is reported in Table B.2. Among the 133 raters, 68 are male, 58 are female; seven raters did not report their gender. The mean age of the 133 raters is 38.42 (SD = 11.84). The raters are from 33 states and the District of Columbia in the continental United States. For birthplace and current residence, 12 raters only responded as "usa" without specifiying the city and state.

Table B.2: Raters' Current Residences and Birthplaces (Experiment 2)

Current Residences	Numbers (CR)	Birthplace	Numbers (BP)
alabama	2	alabama	1
arizona	1	arizona	2
arkansas	2	arkansas	0
california	15	california	17
colorado	1	colorado	3
delaware	1	delaware	0
district of columbia	0	district of columbia	2
florida	7	florida	2
georgia	1	georgia	0
idaho	3	idaho	1
illinois	4	illinois	4
indiana	4	indiana	2
iowa	4	iowa	2
kansas	1	kansas	1
kentucky	0	kentucky	1
maryland	0	maryland	1
michigan	8	michigan	8
minnesota	2	minnesota	1
mississippi	0	mississippi	1
nevada	0	nevada	3
new hampshire	0	new hampshire	1
nebraska	1	nebraska	0
new jersey	8	new jersey	8
new york	11	new york	13
north carolina	1	north carolina	2
ohio	2	ohio	5

oklahoma	1	oklahoma	1
oregon	1	oregon	0
pennsylvania	7	pennsylvania	7
rhode island	1	rhode island	3
south carolina	1	south carolina	3
south dakota	0	south dakota	1
tennessee	1	tennessee	2
texas	8	texas	8
virginia	15	virginia	8
washington	4	washington	4
west virginia	2	west virginia	2
wisconsin	1	wisconsin	1
usa	12	usa	12

Appendix C: Mean Ratings Across the U.S.

Figures C.1 and C.2 illustrates the arithmetic mean ratings of the 100 audio stimuli by raters from each state. The darker the color the higher the mean ratings (i.e., more accented). For example, raters from Nevada judged the same 100 stimuli as being more accented than did raters from Illinois.

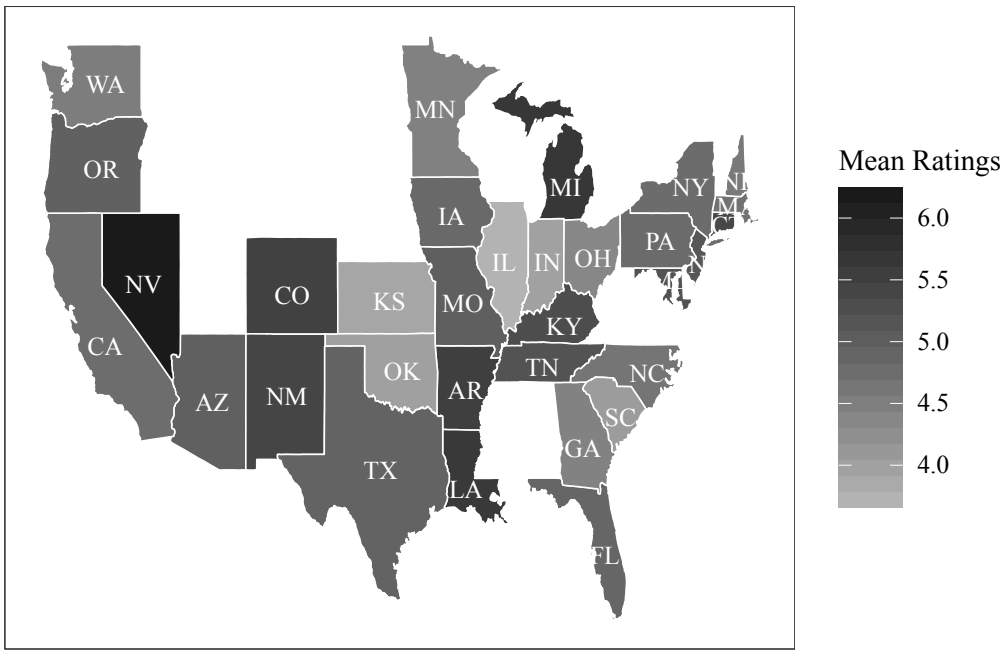

Figure C.1: Mean Ratings by Birthplace (Experiment 1)

Figure C.2: Mean Ratings by Birthplace (Experiment 2)

Bibliography

Adami, Andre G., Radu Mihaescu, Douglas A. Reynolds, and John J. Godfrey (2003). Modeling prosodic dynamics for speaker recognition. *Proceedings.(ICASSP'03)*. Vol. 4. 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. IV–788.

Altendorf, Ulrike and Dominic Watt (2008). The dialects in the South of England: Phonology. *Varieties of English: The British Isles*. Ed. by Bernd Kortmann and Clive Upton. Mouton de Gruyter, pp. 194–222.

Altmann, Christian F., Maiko Uesaki, Kentaro Ono, Masao Matsuhashi, Tatsuya Mima, and Hidenao Fukuyama (2014). Categorical speech perception during active discrimination of consonants and vowels. *Neuropsychologia* 64, pp. 13–23.

Anderson-Hsieh, Janet, Ruth Johnson, and Kenneth Koehler (1992). The Relationship Between Native Speaker Judgments of Nonnative Pronunciation and Deviance in Segmentais, Prosody, and Syllable Structure. *Language Learning* 42.4, pp. 529–555.

Baayen, R. Harald (2011). Corpus linguistics and naive discriminative learning. *Revista Brasileira de Linguística Aplicada* 11.2, pp. 295–328.

Baayen, R. Harald, Cyrus Shaoul, Jon Willits, and Michael Ramscar (2016). Comprehension without segmentation: A proof of concept with naive discriminative learning. *Language, cognition and neuroscience* 31.1, pp. 106–128.

Bates, Douglas, Martin Maechler, Ben Bolker, and Steven Walker (2014). lme4: Linear mixed-effects models using Eigen and S4. *R package version* 1.7, pp. 1–23.

Berman, Marc G., John Jonides, and Richard L. Lewis (2009). In search of decay in verbal short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 35.2, p. 317.

Best, Catherine T. (1995). A direct realist view of cross-language speech rerception. *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*. Ed. by Winifred Strange. York Press., pp. 171–204.

Boersma, Paul and David Weenink (2015). *Praat: Doing phonetics by computer [Computer program]. Version 5.3. 23. Retrieved January 24, 2015*.

Bonatti, Luca L., Marcela Pena, Marina Nespor, and Jacques Mehler (2005). Linguistic constraints on statistical computations: The role of consonants and vowels in continuous speech processing. *Psychological Science* 16.6, pp. 451–459.

Braun, Bettina, Kristin Lemhöfer, and Nivedita Mani (2011). Perceiving unstressed vowels in foreign-accented English. *The Journal of the Acoustical Society of America* 129.1, pp. 376–387.

Campbell, Heather, Daphna Harel, Elaine Hitchcock, and Tara McAllister Byun (2018). Selecting an acoustic correlate for automated measurement of American English rhotic production in children. *International Journal of Speech-Language Pathology* 20.6, pp. 635–643.

Carignan, Christopher (2017). Covariation of nasalization, tongue height, and breathiness in the realization of F1 of Southern French nasal vowels. *Journal of Phonetics* 63, pp. 87–105.

Chamorro, Gloria, Antonella Sorace, and Patrick Sturt (2016). What is the source of L1 attrition? The effect of recent L1 re-exposure on Spanish speakers under L1 attrition. *Bilingualism: Language and Cognition* 19.3, pp. 520–532.

Chan, Kit Ying, Michael D. Hall, and Ashley A. Assgari (2016). The role of vowel formant frequencies and duration in the perception of foreign accent. *Journal of Cognitive Psychology* 29.1, pp. 1–12.

Chodroff, Eleanor and Colin Wilson (2017). Structure in talker-specific phonetic realization: Covariation of stop consonant VOT in American English. *Journal of Phonetics* 61, pp. 30–47.

Chomsky, Noam (1995). *The Minimalist Program*. MIT Press. 436 pp.

Clarke, Sandra (2008). Newfoundland English: Phonology. *Varieties of English* 2, pp. 161–180.

Cutler, Anne, Nuria Sebastián-Gallés, Olga Soler-Vilageliu, and Brit Van Ooijen (2000). Constraints of vowels and consonants on lexical selection: Cross-linguistic comparisons. *Memory & Cognition* 28.5, pp. 746–755.

Danks, David (2003). Equilibria of the Rescorla–Wagner model. *Journal of Mathematical Psychology* 47.2, pp. 109–121.

De Jong, Nivja H. and Ton Wempe (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods* 41.2, pp. 385–390.

Demuth, Katherine, Jennifer Culbertson, and Jennifer Alter (2006). Word-minimality, epenthesis and coda licensing in the early acquisition of English. *Language and Speech* 49.2, pp. 137–173.

Derwing, Tracey M. and Murray J. Munro (1997). Accent, intelligibility, and comprehensibility. *Studies in second language acquisition* 19.1, pp. 1–16.

Di Paolo, Marianna and Alice Faber (1990). Phonation differences and the phonetic content of the tense-lax contrast in Utah English. *Language Variation and Change* 2.2, pp. 155–204.

Difallah, Djellel, Elena Filatova, and Panos Ipeirotis (2018). Demographics and Dynamics of Mechanical Turk Workers. *Proceedings of the 18th ACM International Conference on Web Search and Data Mining (WSDM)*, pp. 135–143.

Dorman, Michael F., Lawrence J. Raphael, and David Isenberg (1980). Acoustic cues for a fricative-affricate contrast in word-final position. *Journal of Phonetics* 8.4, pp. 397–405.

Dupoux, Emmanuel, Kazuhiko Kakehi, Yuki Hirose, Christophe Pallier, and Jacques Mehler (1999). Epenthetic vowels in Japanese: A perceptual illusion? *Journal of Experimental Psychology: Human Perception and Performance* 25.6, pp. 1568–1578.

Edwards, Jette G. Hansen (2011). Deletion of /t, d/ and the Acquisition of Linguistic Variation by Second Language Learners of English. *Language Learning* 61.4, pp. 1256–1301.

Enochson, Kelly and Jennifer Culbertson (2015). Collecting psycholinguistic response time data using Amazon Mechanical Turk. *PloS one* 10.3, e0116946.

Fisher, Ronald Aylmer and Frank Yates (1963). *Statistical tables for biological, agricultural and medical research*. 6th ed. Oliver and Boyd.

Flege, James Emil (1984). The detection of French accent by American listeners. *The Journal of the Acoustical Society of America* 76.3, pp. 692–707.

Flege, James Emil and Kathryn L. Fletcher (1992). Talker and listener effects on degree of perceived foreign accent. *The Journal of the Acoustical Society of America* 91.1, pp. 370–389.

Flege, James Emil and Murray J. Munro (1994). The word unit in second language speech production and perception. *Studies in Second Language Acquisition* 16.4, pp. 381–411.

Floccia, Caroline, Joseph Butler, Frédérique Girard, and Jeremy Goslin (2009). Categorization of regional and foreign accent in 5-to 7-year-old British children. *International Journal of Behavioral Development* 33.4, pp. 366–375.

Frost, Wende (2013). Shibboleth: An Automated Foreign Accent Identification Program. (Doctoral dissertation). Arizona State University.

Garellek, Marc (2019). The phonetics of voice. *The Routledge Handbook of Phonetics*. Ed. by William F. Katz and Peter F. Assmann. Routledge.

Giorgino, Toni (2009). Computing and visualizing dynamic time warping alignments in R: the dtw package. *Journal of statistical Software* 31.7, pp. 1–24.

Gluszek, Agata and John F. Dovidio (2010). Speaking with a nonnative accent: Perceptions of bias, communication difficulties, and belonging in the United States. *Journal of Language and Social Psychology* 29.2, pp. 224–234.

Goldinger, Stephen D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review* 105.2, pp. 251–279.

González-Bueno, Manuela (1997). Voice-onset-time in the perception of foreign accent by native listeners of Spanish. *International Review of Applied Linguistics in Language Teaching* 35.4, pp. 251–268.

Gordon, Matthew J. (2008). New York, Philadelphia, and other northern cities: Phonology. *Varieties of English* 2, pp. 67–86.

Gouskova, Maria and Nancy Hall (2009). Acoustics of epenthetic vowels in Lebanese Arabic. *Phonological argumentation*. Ed. by Steve Parker. Equinox Publishing Ltd, pp. 203–225.

Grant, Linda and Donna Brinton (2014). *Pronunciation myths: Applying second language research to classroom teaching*. University of Michigan Press.

Green, Lisa J. (2002). *African American English: a linguistic introduction*. Cambridge University Press.

Gu, Chong (2014). Smoothing spline ANOVA models. *Journal of Statistical Software* 58.5, pp. 1–25.

Guy, Gregory R. (1991). Explanation in variable phonology: An exponential model of morphological constraints. *Language Variation and Change* 3.1, pp. 1–22.

Hahn, Laura D. (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. *TESOL Quarterly* 38.2, pp. 201–223.

Hall, Nancy Elizabeth (2003). Gestures and segments: Vowel intrusion as overlap. (Doctoral dissertation). University of Massachusetts Amherst.

— (2011). Vowel Epenthesis. *The Blackwell Companion to Phonology*. Ed. by Elizabeth V. Hume Keren Rice Marc van Oostendorp Colin J. Ewen. Wiley-Blackwell, pp. 1576–1596.

Hansen, Jette G. (2004). Developmental sequences in the acquisition of English L2 syllable codas: A preliminary study. *Studies in Second Language Acquisition* 26.1, pp. 85–124.

Hawkins, Sarah and Kenneth N. Stevens (1985). Acoustic and perceptual correlates of the non-nasal–nasal distinction for vowels. *The Journal of the Acoustical Society of America* 77.4, pp. 1560–1575.

Hayes, Bruce and Colin Wilson (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39.3, pp. 379–440.

Hillenbrand, James, Laura A. Getty, Michael J. Clark, and Kimberlee Wheeler (1, 1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America* 97.5, pp. 3099–3111.

Holt, Lori L. and Andrew J. Lotto (2010). Speech perception as categorization. *Attention, Perception, & Psychophysics* 72.5, pp. 1218–1227.

Horvath, Barbara M. (2008). Australian English: Phonology. *Varieties of English* 3, pp. 89–110.

Jilka, Matthias (2000). The contribution of intonation to the perception of foreign accent: Identifying intonational deviations by means of F0 generation and resynthesis. (Doctoral dissertation). University of Stuttgart.

Johnson, Keith (2004a). Aligning phonetic transcriptions with their citation forms. *Acoustics Research Letters Online* 5.2, pp. 19–24.

Johnson, Keith (2004b). Massive reduction in conversational American English. *Spontaneous speech: Data and analysis. Proceedings of the 1st session of the 10th international symposium*. Citeseer, pp. 29–54.

Jongman, Allard (1989). Duration of frication noise required for identification of English fricatives. *The Journal of the Acoustical Society of America* 85.4, pp. 1718–1725.

Jongman, Allard, Ratree Wayland, and Serena Wong (2000). Acoustic characteristics of English fricatives. *The Journal of the Acoustical Society of America* 108.3, pp. 1252–1263.

Kang, Okim, D. O. N. Rubin, and Lucy Pickering (2010). Suprasegmental measures of accentedness and judgments of language learner proficiency in oral English. *The Modern Language Journal* 94.4, pp. 554–566.

Kawahara, Hideki, Toru Takahashi, Masanori Morise, and Hideki Banno (2009). Development of exploratory research tools based on TANDEM-STRAIGHT. *Proceedings: APSIPA ASC 2009: Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference*. Asia-Pacific Signal, Information Processing Association, 2009 Annual Summit, and Conference, International Organizing Committee, pp. 111–120.

Kennedy, Sara and Pavel Trofimovich (2008). Intelligibility, comprehensibility, and accentedness of L2 speech: The role of listener experience and semantic context. *Canadian Modern Language Review* 64.3, pp. 459–489.

Keshet, Joseph, Morgan Sonderegger, and Thea Knowles (2014). *AutoVOT: A tool for automatic measurement of voice onset time using discriminative structured prediction [Computer program]. Version 0.91*.

Klein, Michael (2011). Aspir (at) ing to speak like a native: Tracking voice onset time in the acquisition of English stops. *George Mason University Working Paper in Linguistics* 4.4, pp. 131–136.

Kondrak, Grzegorz (2003). Phonetic alignment and similarity. *Computers and the Humanities* 37.3, pp. 273–291.

Kronrod, Yakov, Emily Coppess, and Naomi H. Feldman (2012). A unified model of categorical effects in consonant and vowel perception. *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, pp. 629–634.

Kuhl, Patricia K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics* 50.2, pp. 93–107.

Labov, William (1997). Resyllabification. *Variation, Change and Phonological Theory*. Red. by Frans L Hinskens, Roeland van Hout, and W. Leo Wetzels. Benjamins, pp. 145–179.

Labov, William, Sharon Ash, and Charles Boberg (2005). *The atlas of North American English: Phonetics, phonology and sound change*. Walter de Gruyter.

Ladefoged, Peter and Ian Maddieson (1996). *Sounds of the Worlds Languages*. 1st ed. Wiley-Blackwell.

Liberman, Alvin M., Katherine Safford Harris, Howard S. Hoffman, and Belver C. Griffith (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology* 54.5, p. 358.

Lippi-Green, Rosina (2012). *English with an Accent: Language, Ideology, and Discrimination in the United States*. 2nd ed. Routledge.

Lisker, Leigh (1986). "Voicing" in English: A catalogue of acoustic features signaling/b/versus/p/in trochees. *Language and speech* 29.1, pp. 3–11.

Lotto, Andrew J., Lori L. Holt, and Kieth R. Kluender (1997). Effect of voice quality on perceived height of English vowels. *Phonetica* 54.2, pp. 76–93.

Magen, Harriet S. (1998). The perception of foreign-accented speech. *Journal of Phonetics* 26.4, pp. 381–400.

Major, Roy C. (1986). Paragoge and degree of foreign accent in Brazilian English. *Second Language Research* 2.1, pp. 53–71.

— (1987). Phonological similarity, markedness, and rate of L2 acquisition. *Studies in Second Language Acquisition* 9.1, pp. 63–82.

— (2012). Foreign Accent. *The Encyclopedia of Applied Linguistics*. Ed. by Carol A. Chapelle. 1st ed. John Wiley.

Mareüil, Philippe Boula de and Bianca Vieru-Dimulescu (2006). The Contribution of Prosody to the Perception of Foreign Accent. *Phonetica* 63.4, pp. 247–267.

Mazzoni, D. and R. Dannenberg (2000). *Audacity [software]. Pittsburg*. PA: Carnegie Mellon University.

McAuliffe, Michael, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. *Interspeech*, pp. 498–502.

McCullough, Elizabeth A. (2013). Acoustic correlates of perceived foreign accent in non-native English. (Doctoral dissertation). The Ohio State University.

McDermott, Wendy Lee Calla (1986). The Scalability of Degrees of Foreign Accent. (Doctoral dissertation). Cornell University.

Miller, George A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* 63.2, p. 81.

Minematsu, Nobuaki, Shun Kasahara, Takehiko Makino, Daisuke Saito, and Keikichi Hirose (2014). Speaker-basis Accent Clustering Using Invariant Structure Analysis and the Speech Accent Archive. *Odyssey: The Speaker and Language Recognition Workshop*. ISCA tutorial and research workshop of Odyssey, pp. 158–165.

Mirman, Daniel, Lori L. Holt, and James L. McClelland (2004). Categorization and discrimination of nonspeech sounds: Differences between steady-state and rapidly-changing acoustic cues. *The Journal of the Acoustical Society of America* 116.2, pp. 1198–1207.

Morrill, Tuuli (2015). Implementation of Phrasal Prosody by Native and Non-native Speakers of English: SS ANOVA for Multi-syllabic Intonation Contours. *ICPHS Proceedings*.

Moulines, Eric and Francis Charpentier (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication* 9.5, pp. 453–467.

Munro, Murray J. (1993). Productions of English Vowels by Native Speakers of Arabic: Acoustic Measurements and Accentedness Ratings. *Language and Speech* 36.1, pp. 39–66.

Munro, Murray J. and Tracey M. Derwing (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning* 45.1, pp. 73–97.

Munro, Murray J. and Tracey M. Derwing (1998). The Effects of Speaking Rate on Listener Evaluations of Native and Foreign-Accented Speech. *Language Learning* 48.2, pp. 159–182.

— (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech the role of speaking rate. *Studies in Second Language Acquisition* 23.4, pp. 451–468.

— (2006). The functional load principle in ESL pronunciation instruction: An exploratory study. *System* 34.4, pp. 520–531.

Munro, Murray J, Tracey M Derwing, and James E Flege (1999). Canadians in Alabama: a perceptual study of dialect acquisition in adults. *Journal of Phonetics* 27.4, pp. 385–403.

Nerbonne, John, Wilbert Heeringa, Erik Van den Hout, Peter Van der Kooi, Simone Otten, and Willem Van de Vis (1996). Phonetic distance between Dutch dialects. *CLIN VI: proceedings of the sixth CLIN meeting*, pp. 185–202.

Nespor, Marina, Marcela Peña, and Jacques Mehler (2003). On the different roles of vowels and consonants in speech processing and language acquisition. *Lingue e linguaggio* 2.2, pp. 203–230.

Niebuhr, Oliver (2017). On the perception of "segmental intonation": F0 context effects on sibilant identification in German. *EURASIP Journal on Audio, Speech, and Music Processing* 2017.1, p. 19.

Nielsen, Kuniko (2011). Specificity and abstractness of VOT imitation. *Journal of Phonetics* 39.2, pp. 132–142.

Ohala, Manjari and John Ohala (2001). Acoustic VC transitions correlate with degree of perceptual confusion of place contrast in Hindi. *Travaux du cercle Linguistique de Copenhague* 31, pp. 265–284.

Park, Hanyong (2013). Detecting foreign accent in monosyllables: The role of L1 phonotactics. *Journal of Phonetics* 41.2, pp. 78–87.

Peterson, Gordon E. and Harold L. Barney (1952). Control methods used in a study of the vowels. *The Journal of the acoustical society of America* 24.2, pp. 175–184.

Pisoni, David B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception & Psychophysics* 13.2, pp. 253–260.

Repp, Bruno H. (1984). Categorical perception: Issues, methods, findings. *Speech and Language: Advances in Basic Research and Practice*. Ed. by Norman J. Lass. Vol. 10. Academic Press, pp. 243–335.

Rescorla, Robert A. and Allan R. Wagner (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current research and theory*. Ed. by William Frederick Prokasy Abraham H. Black. Vol. 2. Appleton-Century-Crofts, pp. 64–99.

Rilliard, Albert, Alexandre Allauzen, and Philippe Boula de Mareüil (2011). Using Dynamic Time Warping to Compute Prosodic Similarity Measures. *INTERSPEECH*, pp. 2021–2024.

Riney, Timothy J. and James E. Flege (1998). Changes over time in global foreign accent and liquid identifiability and accuracy. *Studies in Second Language Acquisition* 20.2, pp. 213–243.

Riney, Timothy J., Mari Takada, and Mitsuhiko Ota (2000). Segmentals and global foreign accent: The Japanese flap in EFL. *Tesol Quarterly* 34.4, pp. 711–737.

Riney, Timothy J. and Naoyuki Takagi (1999). Global foreign accent and voice onset time among Japanese EFL speakers. *Language Learning* 49.2, pp. 275–302.

Riney, Timothy J., Naoyuki Takagi, and Kumiko Inutsuka (2005). Phonetic parameters and perceptual judgments of accent in English by American and Japanese listeners. *Tesol Quarterly* 39.3, pp. 441–466.

Rognoni, Luca and Maria Grazia Busà (2014). Testing the Effects of Segmental and Suprasegmental Phonetic Cues in Foreign Accent Rating: An Experiment Using Prosody Transplantation. *Proceedings of the International Symposium on the Acquisition of Second Language Speech Concordia Working Papers in Applied Linguistics*. Vol. 5, pp. 547–560.

Romberg, Alexa R. and Jenny R. Saffran (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews. Cognitive Science* 1.6, pp. 906–914.

Sato, Charlene J. (1984). Phonological processes in second language acquisition: Another look at interlanguage syllable structure. *Language Learning* 34.4, pp. 43–58.

Schaden, Stefan (2006). Evaluation of Automatically Generated Transcriptions of Non-Native Pronunciations using a Phonetic Distance Measure. *LREC*. Citeseer, pp. 2441–2446.

Sharpe, Victoria, Daniel Fogerty, and Dirk-Bart Den Ouden (2015). The role of fundamental frequency and temporal envelope in processing sentences with temporary syntactic ambiguities. *Proceedings of Meetings on Acoustics*. Vol. 21. Acoustical Society of America, p. 060006.

Shen, Han-Ping, Nobuaki Minematsu, Takehiko Makino, Steven H. Weinberger, Teeraphon Pongkittiphan, and Chung-Hsien Wu (2013). Speaker-based accented English clustering using a world English archive. *Speech and Language Technology in Education*. International Speech Communication Association, pp. 184–188.

Shue, Yen-Liang, Patricia Keating, Chad Vicenik, and K. VoiceSauce Yu (2011). A program for voice analysis. *Proceedings of the Seventeenth International Congress of Phonetic Sciences*, pp. 1846–1849.

Sidaras, Sabrina K., Jessica E. D. Alexander, and Lynne C. Nygaard (2009). Perceptual learning of systematic variation in Spanish-accented speech. *The Journal of the Acoustical Society of America* 125.5, pp. 3306–3316.

Sjölander, K (2004). *The Snack Sound Toolkit*. Version 2.2.10.

Solon, Megan (2015). L2 Spanish/l: The Roles of F2 and Segmental Duration in Foreign Accent Perception. *Selected Proceedings of the 6th Conference on Laboratory Approaches to Romance Phonology*, pp. 83–94.

Sprouse, Jon (2010). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods* 43.1, pp. 155–167.

Thomson, R. (2014). Myth 6: Accent reduction and pronunciation instruction are the same thing. *Pronunciation Myths: Applying Second Language Research to Classroom Teaching*. Ed. by Linda Grant and Donna Brinton. University of Michigan Press, pp. 160–187.

van den Doel, Rias (2006). How friendly are the natives? An evaluation of native speaker judgements of foreign-accented British and American English. (Doctoral dissertation). University of Utrecht.

van Ooijen, Brit (1996). Vowel mutability and lexical selection in English: Evidence from a word reconstruction task. *Memory & Cognition* 24.5, pp. 573–583.

Vitevitch, Michael S. and Paul A. Luce (2004). A web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments, & Computers* 36.3, pp. 481–487.

Waniek-Klimczak, Ewa, Arkadiusz Rojczyk, and Andrzej Porzuczek (2015). 'Polglish'in Polish Eyes: What English Studies Majors Think About Their Pronunciation in English. *Teaching and Researching the Pronunciation of English*. Springer, pp. 23–34.

Wayland, Ratree (1997). Non-native Production of Thai: Acoustic Measurements and Accentedness Ratings. *Applied Linguistics* 18.3, pp. 345–373.

Weinberger, Steven, Stephen Kunath, Jill Nelson, and Zhiyan Gao (2019). Crowdsourcing L2 Phonetics. *The 11th annual Pronunciation in Second Language Learning and Teaching*.

Weinberger, Steven H. (2019). *Speech accent archive*. Geroge Mason University. URL: `http://accent.gmu.edu`.

Weinberger, Steven H. and Stephen A. Kunath (2011). The Speech Accent Archive: towards a typology of English accents. *Corpus-based Studies in Language Use, Language Learning, and Language Documentation*. Brill Rodopi, pp. 265–281.

Wells, John C. (1982). *Accents of English*. Vol. 1. Cambridge University Press.

Wester, Mirjam and Cassie Mayo (2014). Accent rating by native and non-native listeners. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 7699–7703.

Whalen, Douglas H. and Andrea G. Levitt (1995). The universality of intrinsic F 0 of vowels. *Journal of Phonetics* 23.3, pp. 349–366.

White, Laurence and Sven L. Mattys (2007). Calibrating rhythm: First language and second language studies. *Journal of Phonetics* 35.4, pp. 501–522.

Wieling, Martijn, Jelke Bloem, Kaitlin Mignella, Mona Timmermeister, and John Nerbonne (2014a). Measuring foreign accent strength in English: Validating Levenshtein distance as a measure. *Language Dynamics and Change* 4.2, pp. 253–269.

Wieling, Martijn, John Nerbonne, Jelke Bloem, Charlotte Gooskens, Wilbert Heeringa, and R. Harald Baayen (2014b). A cognitively grounded measure of pronunciation distance. *PloS one* 9.1, e75734.

Woods, Kevin J.P., Max Siegel, James Traer, and Josh H. McDermott (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception & Psychophysics* 79.7, pp. 2064–2072.

Yoon, Kyuchul (2007). Imposing native speakers' prosody on non-native speakers' utterances: The technique of cloning prosody. *Journal of the Modern British & American Language & Literature* 25.4, pp. 197–215.

Yuan, Jiahong and Mark Liberman (2008). Speaker identification on the SCOTUS corpus. *Journal of the Acoustical Society of America* 123.5, p. 3878.

Zielinski, Beth W. (2008). The listener: No longer the silent partner in reduced intelligibility. *System* 36.1, pp. 69–84.

Curriculum Vitae

Zhiyan Gao graduated from Soochow University (Suzhou, China) in 2009 with a BA in Chinese and a BA in Economics. He received an MA in English with a concentration in Linguistics from George Mason University in 2012. Zhiyan was the recipient of the 2013 GMU Presidential Scholarship, providing three years of tuition and assistantship.