Influence of Poor Fit Vowels on Perception of Consonants

Yuka Muratani

A thesis submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Arts in the Department of Linguistics

Chapel Hill
2017

Approved by:

A. Elliott Moreton

Jennifer L. Smith

Katya Pertsova

ABSTRACT

Yuka Muratani: Influence of Poor Fit Vowels on Perception of Consonants
(Under the direction of A. Elliott Moreton)

The present study investigated native English listeners' perception of an ambiguous fricative noise from a [s]-[ʃ] continuum followed/preceded by a poor fit vowel—either one of the [i]s that have higher/lower formant frequencies than a good exemplar of English [i], or [u]s that have higher/lower formant frequencies than a good exemplar of English [u]. The main questions that the present study intended to address were, i) whether listeners would show perceptual contextual dissimilation (a.k.a. compensation for coarticulation, Mann & Repp, 1980, 1981) or listeners would show perceptual contextual assimilation (a.k.a. parsing, Fowler, 1984); and ii) whether listeners would respond to the stimuli according to their phonological analysis of the segments (Kingston et al., 2011) or according to the actual phonetic details of the segments (Whalen, 1989). The results were that the listeners showed perceptual contextual dissimilation for their broad (more abstract) phonological categorization of [i] and [u]. However, when the listeners were sensitive to the phonetic details of the segments, the listeners showed perceptual contextual assimilation. The listeners somehow, however, were not sensitive to the phonetic details of poor fit vowels when the stimuli were identified as [si] and [ʃi]. Although it is hard to come to a solid conclusion from these response patterns, the results at least indicate that listeners may be able to parse vowels using their native language knowledge, and dynamically adjust the acoustic discrepancy by showing perceptual contextual assimilation.

keywords: speech perception, compensation for coarticulation, parsing, phonetic details

To my fiancé, Masayuki.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my thesis advisor, A. Elliott Moreton, for having *so many* meetings with me and steering me in the right direction. This thesis is a direct result of the "Second Language Phonetics/Phonology" class taught by him when I first started my graduate studies here at UNC, Chapel Hill. I will miss visiting his office. I would also like to express my gratitude to the rest of my committee, Jennifer L. Smith and Katya Pertsova, for their insightful comments and suggestions on this thesis, as well as hard questions in our meetings. I would like to thank Chris Wiesen at the Odum Institute for his statistical analysis and advice. I would also like to thank the members in P-side their comments on the stimuli, and my colleagues in our linguistics department for participating in my pilot experiments. Particularly, I would like to thank Emily Moeng, Anissa Neal, and her sister Leah Neal for listening to the stimuli and judging them again and again. I would also like to thank my best friend, Trevor McCarley, for correcting grammatical errors. Finally, this study was supported by The Juanita A. Melchert Graduate Student Excellence Fund. I could not have finished this study without the fund.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1. Introduction

In a normal speech production, when a continuous string of phonetic segments is produced, each segment overlaps with its neighboring segments, "which results in segments generally appearing assimilated to their contexts" (Keating, 1990). This is because there is only one single vocal tract for various segments, and thus, the single vocal tract needs to change its shape for each segment dynamically. The results are gestural overlaps. For example, when the segment [s] is coproduced with the segment [i] in the English word 'see' ([si]), the tongue body and tip move forwards before the fricative-vowel boundary because the following segment [i] is articulated with the front portion of the oral cavity (Katz & Bharadwaj, 2001). When the segment [s] is coproduced with the segment [u] in the English word 'sue' ([su]), on the other hand, the tongue body and tip move backwards before the fricative-vowel boundary because the following segment [u] is articulated with the back portion of the oral cavity (Katz & Bharadwaj, 2001). In addition, when the segment [s] is coproduced with the segment [u], the lip rounding gesture for the segment [u] starts before the fricative-vowel boundary (Bell-Berti & Harris, 1979). These gestural overlaps in speech production are referred to as *coarticulation*.

A consequence of coarticulation is known to be "lack of invariance" in acoustic information (e.g., Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Liberman, Delattre, Cooper, & Gerstman, 1954; Liberman & Mattingly, 1985). For example, the coarticulation of the segment [s] and the segment [i]—which leads the preceding fricative segment to be articulated with the relatively forward portion of the palate—would cause main fricative spectral peaks produced at relatively higher frequencies. Likewise, the coarticulation of the segment [s] and the segment [u]—which leads the preceding fricative segment to be articulated with the relatively backward portion of the palate, and with some degree of anticipatory lip rounding—cause main fricative spectral peaks produced at relatively lower frequencies (Soli, 1981). The changes could range from 100 to 200 Hz (Soli, 1981). The variance of acoustic information seems to pose many problems to listeners at first glance; however, scholars (Elman & McClelland, 1986) claim coarticulation is rather lawful and beneficial to listeners.

The question is, how do listeners solve this "lack of invariance" problem in speech perception so successfully? There are various empirical studies that show listeners take account of coarticulatory (context) effects and make dynamic adjustments in speech perception. The present study is concerned with 2 types of such speech perception mechanisms, which are referred to as *perceptual contextual dissimilation* and *perceptual contextual assimilation* in this thesis.

In perceptual contextual dissimilation—which is more known as "compensation for coarticulation" (Mann & Repp, 1980, 1981), listeners hear an ambiguous target segment as having relatively high spectral frequencies if the context segment has relatively low spectral frequencies. If the context segment has relatively high spectral frequencies, on the other hand, listeners hear an ambiguous target segment as having relatively low spectral frequencies. That is, in perceptual contextual dissimilation, listeners hear the target segment as different from its context segment in terms of the spectral frequencies of the interacting segments (relevant studies will be reviewed in the upcoming chapter). For example, if an ambiguous segment around the phonetic category boundary of [s] and [ʃ] is surrounded by the vowel [u] as its context segment, since the vowel [u] has relatively low formant frequencies, listeners tend to identify an ambiguous noise as [s], which has its main spectral peaks at relatively high frequencies, rather than [ʃ], which has its main spectral peaks at relatively low frequencies. Figure 1 illustrates this type of speech perception.

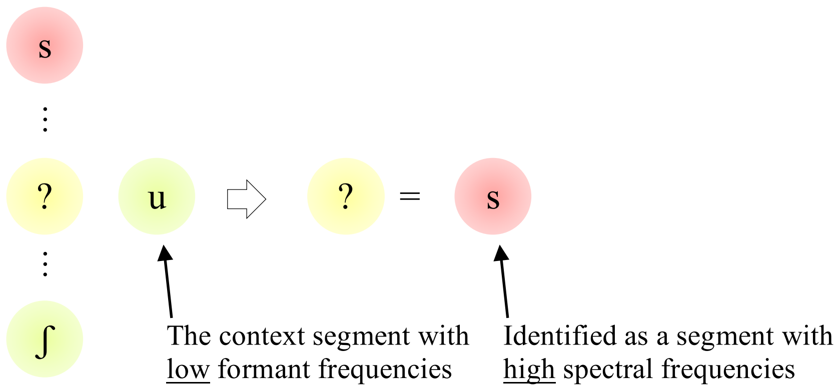
s
?
ʃ
u
?
=
s
The context segment with low formant frequencies
Identified as a segment with high spectral frequencies

Figure 1. Illustration of perceptual contextual dissimilation

On the other hand, in perceptual contextual assimilation—which can be thought of as "parsing" (Fowler, 1984, 1996; Fowler & Brown, 2000; Fowler & Smith, 1986), listeners detect possible coarticulation information in the context segment and attribute it to the target segment. Hence, in perceptual contextual assimilation, listeners hear the target segment as similar to its context segment in terms of the spectral frequencies of the interacting segments (studies that observed this type of speech perception will be reviewed in the upcoming chapter). For example, as Figure 2 illustrates, if an ambiguous segment between [s] and [ʃ] is

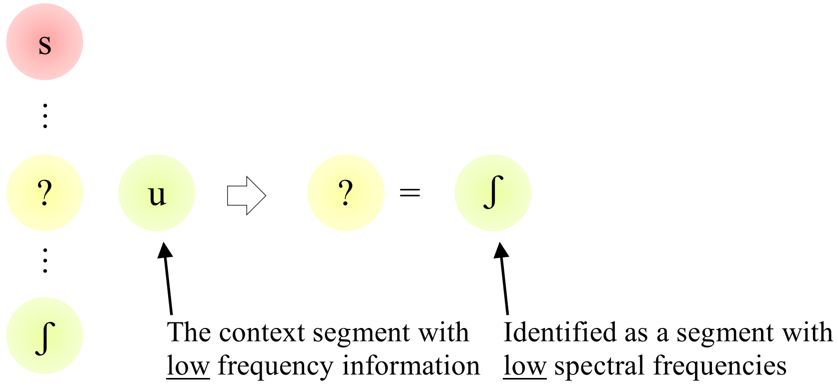
s
?
ʃ
u
?
=
ʃ
The context segment with low frequency information
Identified as a segment with low spectral frequencies

Figure 2. Illustration of perceptual contextual assimilation

surrounded by the vowel [u] as its context segment, and if the vowel [u] has some coarticulation information that indicates low spectral frequencies, listeners tend to identify an ambiguous noise as [ʃ], which has its main spectral peaks at relatively low frequencies, rather than [s], which has its main spectral peaks at relatively high frequencies.

The motivation of the present study is Whalen (1989) and Kingston, Kawahara, Mash, & Chambless (2011), which will be reviewed in the upcoming chapter. Briefly speaking, both of these studies investigated English listeners' perception of ambiguous consonants in the context of non-native-like (poor fit) vowels; however, the results (or explanations) provided by these studies were somewhat paradoxical. According to Whalen, the English listeners showed perceptual contextual dissimilation according to the actual phonetic details of the segments even though the segments were non-native-like. However, the English listeners in Kingston et al.'s study did not show the same pattern. According to Kingston et al., the English listeners showed perceptual contextual dissimilation according to their phonological analysis of the segments. The issue here is that Whalen's claim that listeners would show perceptual contextual dissimilation according to the phonetic details of the segments cannot explain the English listeners' response patterns observed in Kingston et al.'s study. Likewise, Kingston at al.'s explanation cannot account for the English listeners' response patterns observed in Whalen's study. For this, I hypothesized that the English listeners in Kingston et al.'s study might have been showing perceptual contextual assimilation, which made the response patterns look like they were showing perceptual contextual dissimilation according to their phonological analysis of the segments. In order to test this hypothesis against Whalen's claim and Kingston et al.'s claim, an experiment was carried out.

There were two research questions that were investigated: First, the present study investigated if listeners would show perceptual contextual dissimilation or perceptual contextual assimilation when an ambiguous fricative noise from a [s]-[ʃ] continuum is followed/preceded by a poor fit vowel (an [i] that has higher formant frequencies than a good exemplar of English [i], an [i] that has lower formant frequencies than a good exemplar of English [i], an [u] that has higher formant frequencies than a good exemplar of English [u], and an [u] that has lower formant frequencies than a good exemplar of English [u]). Second, the present study investigated if listeners would respond to the stimuli according to the phonetic details of the segments (Whalen, 1989) or according to their phonological analysis of the segments (Kingston et al., 2011).

The results were that [ʃ] responses occurred more often with [i] responses than [u] responses. That is, the listeners showed perceptual contextual dissimilation for their broad (more abstract) phonological categorization of [i] and [u]. However, a statistical analysis has shown that when the listeners were sensitive to the phonetic details of the segments, the listeners identified an ambiguous fricative noise as [ʃ] more often for a vowel with lower formant frequencies than for a vowel with higher formant frequencies. This was because the [u] with lower formant frequencies than a good exemplar of English [u] triggered more [ʃ] responses than the [u] with higher formant frequencies than a good exemplar of English [u] for both the CV and VC stimuli. Furthermore, an [i] with lower formant frequencies than a good exemplar of English [i] triggered more [ʃ] responses than an [i] with higher formant frequencies than a good exemplar of English [i] for the VC stimuli. That is, the listeners showed perceptual contextual assimilation according to the phonetic details of the vowels in these conditions. However, when an ambiguous vowel was identified as [i] for the CV stimuli—in other words, when the stimuli were identified as [si] and [ʃɪ]—the listeners were not sensitive to the phonetic differences of the vowels. Although it is hard to come to a solid conclusion from these response patterns, the results at least indicate that listeners may be able to parse vowels using their native language knowledge, and dynamically adjust the acoustic discrepancy by showing perceptual contextual assimilation.

These results are important to the research field of speech perception for several reasons. First, recent studies on perceptual contextual dissimilation often utilize foreign languages, and assume their results are reflecting the listeners' pure perceptual contextual dissimilation. The results in the present study, however, indicate that listeners' response patterns may not reflect pure perceptual contextual dissimilation if foreign vowels are used. Non-native vowels may make listeners show perceptual contextual assimilation, which affects the perception of the neighboring consonant. The same may apply to studies on second language acquisition. Identification of non-native consonants may be affected by perceptual contextual assimilation triggered by the perception of non-native vowels.

Second, the present results add new information to the concept of "parsing." Fowler and her colleagues (Fowler, 1984, 1996; Fowler & Brown, 2000; Fowler & Smith, 1986) have shown that listeners parse acoustic information along with "gestural lines." However, the present results suggest that parsing should not be limited to gestural lines. Listeners seem to be able to parse vowels according to their native knowledge, and attribute the residual spectral frequencies to the neighboring segment.

Lastly, the present results are also important for the theory of sound-change developed by Ohala (1981, 2012). In his theory, sound change is due to "hypocorrection" (listeners' failure of correcting coarticulatory effects) or "hypercorrection" (listeners' unnecessary application of correction), where the "correction" is based on perceptual contextual dissimilation. However, the present results indicate that listeners' "correction" could be via perceptual contextual assimilation as well.

The remainder of this thesis is organized as follows: Chapter 2 provides conceptual background for the present study and the motivation of the present study; Chapter 3 describes research questions, hypotheses, and their predictions; Chapter 4 explains the experimental methods; Chapter 5 presents the statistical analysis and the results; and Chapter 6 discusses the results and their theoretical implications.

CHAPTER 2. BACKGROUND

2.1. Introduction

A consequence of coarticulation is "lack of invariance" in acoustic information. Despite that this may sound problematic to listeners, there are various empirical studies that show listeners take account of coarticulatory (context) effects and make dynamic adjustments in speech perception. The present study is concerned with two types of such speech perception mechanisms: i) *perceptual contextual dissimilation*—which is more known as "compensation for coarticulation" (Mann & Repp, 1980, 1981), and ii) *perceptual contextual assimilation*—which is often referred to as "parsing" (Fowler, 1984, 1996; Fowler & Brown, 2000; Fowler & Smith, 1986). Thus, this chapter will begin by briefly reviewing these two types of speech perception mechanisms (sections 2.2. and 2.3.).

The following section (2.4.) will discuss the motivation of the present study, which are the results from Whalen (1989) and Kingston et al. (2011). Their studies were both on perceptual contextual dissimilation; however, their results were somewhat paradoxical. While Whalen's study suggested that listeners show perceptual contextual dissimilation according to the actual phonetic details of the segments they hear even if the segments are non-native-like, Kingston et al.'s study suggested that listeners show perceptual contextual dissimilation according to their phonological analysis of the segments they hear if the segments are non-native-like. The section will explain why these results are considered to be paradoxical.

The last section (2.5.) will propose a possible solution to this paradox. The section will propose another possible interpretation of the results of Kigston et al.'s study, and another hypothetical concept of "parsing."

2.2. Perceptual Contextual Dissimilation (a.k.a. Compensation for Coarticulation)

The phenomenon of perceptual contextual dissimilation has been reported by various scholars (e.g., Fowler 2006; Kang, Johnson, & Finley, 2016; Kingston et al., 2011; Kunisaki & Fujisaki, 1977; Lotto & Kluender, 1998; Lotto, Kluender, & Holt, 1997; Mann, 1980; Mann & Repp, 1980, 1981; Viswanathan, Magnuson, & Fowler, 2010; Whalen, 1981, 1989) ever since the "lack of invariance" problem was raised. The general findings of these studies are that listeners hear an ambiguous target segment as having relatively high spectral frequencies if the context segment has relatively low spectral frequencies, and hear an ambiguous target segment as having relatively low spectral frequencies if the context segment has relatively high spectral frequencies.

Kunisaki and Fujisaki (1977), for example, investigated how the consequence of coarticulation of a fricative noise and a vowel is reflected in the speech perception of fricative noises. They synthesized CV stimuli where the consonant was an ambiguous fricative noise from a [s]-[ʃ] continuum and the vowel was either [a] or [u], and tested Japanese listeners' perception of the noises. The results were that the listeners identified an ambiguous fricative noise as [s] more often when the context vowel was [u] than when the context vowel was [a]. That is, the phonetic category boundary of [s] and [ʃ] shifted towards [ʃ] when the context vowel was [u]. Mann and Repp (1980) and Whalen (1981) replicated this study with native English listeners, and observed the same kind of response patterns.

As described in the previous chapter, in a natural speech production, when the segment [s] is coarticulated with the segment [u], the lip rounding gesture for the segment [u] begins before the fricative-vowel boundary. The result of this coarticulation is that the preceding fricative [s] is produced with its main spectral peaks at the lower frequencies when it is followed by the rounded vowel [u] than when it is followed by the unrounded vowel [a] (Kunisaki & Fujisaki, 1977). Likewise, when the segment [ʃ] is coarticulated with the segment [u], the preceding fricative [ʃ] is produced with its main spectral peaks at the lower frequencies when it is followed by the rounded vowel [u] than when it is followed by the unrounded vowel [a] (Kunisaki & Fujisaki, 1977). From these observations, Kunisaki and Fujisaki explained that the listeners know the consequence of coarticulation and shifted the phonetic category boundary to the lower end when followed by the vowel [u] so that they can identify fricative noises correctly.

The interpretation as to why such phenomenon happens, however, differs among theorists. "Gesturalists" (e.g., Fowler 2006; Kunisaki & Fujisaki, 1977; Mann, 1980; Mann & Repp, 1980, 1981; Viswanathan et al., 2010) claim that this phenomenon is due to listeners' "compensation for coarticulation" (Mann, 1980; Mann & Repp, 1980, 1981) as described above. On the contrary, "auditorists" (e.g., Kingston et al., 2011; Lotto, Kluender, & Holt, 1997; Lotto & Kluender, 1998; Holt & Lotto, 2002) claim that this phenomenon reflects a general auditory process that makes listeners perceive an ambiguous segment as having relatively low spectral frequencies after a segment with relatively high spectral frequencies (or perceive an ambiguous segment as having relatively high spectral frequencies after a segment with relatively low spectral frequencies). That is, the reason the listeners identified more fricative noises as [s] when the following vowel was [u] than when the following vowel was [a] in Kunisaki and Fujisaki's (1977) study is that since the vowel [u] has relatively low formant frequencies compared to the vowel [a], the preceding ambiguous fricative noise was perceived as having its main spectral peaks at relatively higher spectral frequencies, which is [s]. Some other theorists (e.g., Elman & McClelland, 1988; Pitt & McQueen, 1998) claim that this phenomenon has to do with something other than acoustics and phonetics about coarticulation. These theorists claim that this phenomenon comes from some higher-level sources such as lexical knowledge (Elman & McClelland, 1988) or the knowledge of phonotactic probabilities (Pitt & McQueen, 1998).

Discriminating between these accounts is not at issue in the present study. No matter which account is right, listeners hear the target segment as different from its context. If the context segment has relatively low spectral frequencies, the target segment is identified as a segment with relatively high spectral frequencies. If the context segment has relatively high spectral frequencies, the target segment is identified as a segment with relatively low spectral frequencies. Although this phenomenon is more commonly known as "compensation for coarticulation" or "auditory contrast," in order to be neutral regarding these accounts, this phenomenon will be referred to as perceptual contextual dissimilation in this thesis.

2.3. Perceptual Contextual Assimilation (a.k.a. Parsing)

The other type of speech perception mechanisms that the present study is concerned with is perceptual contextual assimilation. The phenomenon of perceptual contextual assimilation also has been reported by various scholars (e.g., Beddor et al., 2013; Fowler, 1984; Fowler & Smith, 1986; Marslen-Wilson & Warren, 1994; Martin & Bunnell, 1981, 1982; Whalen, 1984, 1991; Warren & Marslen-Wilson, 1987; Yeni-Komshian & Soli, 1981). The general findings of these studies are that listeners detect possible coarticulation information in the context segment and attribute it to (or identify) the target segment.

For example, Yeni-Komshian and Soli (1981) excerpted fricative noises [s], [z], [ʃ], and [ʒ] that were produced before a vowel ([ɑ], [i], or [u]), and presented the fricative noises to native English listeners. The listeners were told that a vowel following a fricative noise originally had been excised, and instructed to identify the vowel that has been excised for each fricative noise. The results were that the listeners could identify excised vowels with better-than-chance accuracy. Particularly, the high vowels, [i] and [u], were more likely to be identified correctly than the vowel [ɑ]. Whalen (1983) conducted a similar study with fricative noises originally preceded by vowels, and obtained the same findings.

As described in the previous chapter, when the segment [s] is coarticulated with the segment [i] in the English word 'see' ([si]), for example, the tongue body and tip move forwards before the fricative-vowel boundary because the following segment [i] is articulated with the front portion of the oral cavity (Katz & Bharadwaj, 2001). This yields a domain *x* to contain information about both segments [s] and [i]. From this, Yeni-Komshian and Soli, as well as Whalen, concluded that the listeners used the coarticulatory information of the excited vowels in fricative noises to identify the excised vowels.

As Fowler and her colleagues (Fowler, 1984; Fowler & Brown, 2000; Fowler & Smith, 1986) argue, there are two possible ways for segmenting. One is that segmentation lines are drawn vertically to the axis of time at the boundary between two segments. This segmentation is similar to the segmentation strategy that linguists use when measuring segments. The other segmentation strategy is that listeners segment along coarticulatory lines. That is, segmentation lines are drawn horizontally to the axis of time.

Fowler and her colleagues (Fowler, 1984; Fowler & Brown, 2000; Fowler & Smith, 1986) have shown that listeners segment along coarticulatory lines. Fowler (1984), for example, obtained natural utterances of [gi] and [gu], and prepared a pair of phonetically identical stop bursts (g_i & g_i) followed by different vowels, and a pair of phonetically different stop bursts (g_i & g_u) followed by different vowels. In the pair of phonetically identical stop bursts, one item was spliced and the other item was cross-spliced (g_ii & g_iu). In the pair of phonetically different stop bursts, both items were spliced (g_ii & g_uu). They presented these to native English listeners in a 4IAX discrimination task, where the English listeners were instructed to pick a pair that has perceptually similar items. The first pair was supposed to be chosen if listeners segment at the boundary between two segments, and the second pair was supposed to be chosen if listeners segment along with coarticulatory lines. The results were that the listeners picked the pair of phonetically different stop bursts (g_ii & g_uu). From this, Fowler (1984) concluded that listeners segment along coarticulatory lines.

More recent studies (e.g., Beddor, McGowan, Boland, Coetzee, & Brasher, 2013; Dahan, Magnuson, Tanenhaus, & Hogan, 2001) utilized eye-tracking technologies and observed that listeners' use of coarticulatory information in segments is dynamic (along with coarticulatory lines). Beddor et al. (2013), for example, investigated English listeners' use of nasalization in a vowel that precedes a nasal consonant using an eye-tracking method. In this study, English listeners were presented with English monosyllabic words (CVC, e.g., bet/bed, and CVNC, e.g., bent/bend) where the vowel was an oral vowel for CVC words, and where the vowel was either a nasal vowel with the earlier onset of nasalization, or a nasal vowel with the later onset of nasalization for CVNC words. The listeners were presented with visual choice of images that correspond to these words, and heard a word while the listeners' eye-movements were tracked simultaneously. The results showed that as soon as the onset of nasalization in the vowel began, the listeners looked at the image of CVNC opposed to the image of CVC.

Unlike perceptual contextual dissimilation, in perceptual contextual assimilation, listeners hear the target segment as similar to its context. If the context segment has low spectral frequency information, the target segment is identified as having low spectral frequencies. If the context segment has high spectral frequency information, the target segment is identified as having high spectral frequencies.

2.4. Motivation of the Present Study

What motivates the present study is the reports from Whalen (1989) and Kingston et al. (2011). As mentioned earlier, their studies were both on perceptual contextual dissimilation; however, their results were somewhat paradoxical. This section will review these studies.

Whalen's study is somewhat similar to Kunisaki and Fujisaki's (1977) study mentioned above. In Experiment 3 of Whalen's study, Whalen synthesized ambiguous fricative noises between [s] and [ʃ] varying in their pole frequencies, and ambiguous vowels between [i] and [u] varying in their formant frequencies, and prepared a series of CV stimuli. Native English listeners were asked to categorize the stimuli into 'see,' 'sue,' 'she,' and 'shoe.' The results were that [s] responses more often occurred with vowels with lower formant frequencies than vowels with higher formant frequencies. [ʃ] responses more often occurred with vowels with higher formant frequencies than vowels with lower formant frequencies. The listeners were also likely to hear an ambiguous vowel as [i] when the fricative had relatively lower spectral frequencies, and as [u] when the fricative had relatively higher spectral frequencies.

More importantly, if two ambiguous vowels were categorized as [u], the [u] with lower formant frequencies induced more [s] responses than the [u] with higher formant frequencies. Likewise, if two ambiguous vowels were categorized as [i], the [i] with higher formant frequencies induced more [ʃ] responses than the [i] with lower formant frequencies. Also, if two ambiguous fricative noises were categorized as [s], the [s] with higher spectral frequencies induced more [u] responses than the [s] with lower spectral frequencies. If two ambiguous fricative noises were categorized as [ʃ], the [ʃ] with lower spectral frequencies induced more [i] responses than the [ʃ] with higher spectral frequencies. In short, the general findings of this study were that the listeners were likely to hear an ambiguous segment as a sound with relatively lower frequencies before/after the sound with relatively higher frequencies, and the listeners were likely to hear an ambiguous segment as a sound with relatively higher frequencies before/after the sound with relatively lower frequencies.

There are two main findings the present study is especially interested in. The first is that the listeners showed perceptual contextual dissimilation even though all of the fricatives and vowels were ambiguous, in other words, non-native-like to the listeners. Past studies (Kunisaki & Fujisaki, 1977; Mann & Repp, 1980; Whalen 1981) have shown that the listeners would show perceptual contextual dissimilation for an ambiguous fricative noise between [s] and [ʃ] followed by a non-ambiguous vowel. Experiment 2 of Whalen's (1989) study has shown that the listeners would show perceptual contextual dissimilation for an ambiguous vowel between [i] and [u] followed by a non-ambiguous fricative noise. Experiment 3 of Whalen's (1989) study just described above has investigated what would happen if the combination of an ambiguous fricative and an ambiguous vowel were to be judged. The experiment resulted with the listeners showing perceptual contextual dissimilation even for the combination of an ambiguous fricative and an ambiguous vowel.

Second is that the listeners showed perceptual contextual dissimilation according to the phonetic details of the sounds they heard. As described above, when two ambiguous vowels were categorized as [u], for example, the [u] with lower formant frequencies induced more [s] responses than the [u] with higher formant frequencies. If the vowel category, not phonetic details, had triggered [s] responses, the [u] with lower formant frequencies and the [u] with higher formant frequencies would have triggered the same amount of [s] responses. The results shown in Whalen's study, thus, indicate that listeners show perceptual contextual dissimilation according to the phonetic details of the sounds they hear.

However, Kingston et al.'s study exhibits a different picture. Kingston et al. investigated both English and Japanese listeners' perception of an ambiguous stop consonant from a [t]-[k] continuum preceded by a vowel ([i], [e], [u], or [o]) synthesized based on formant frequencies of Japanese vowels. The results were that English listeners heard an ambiguous stop consonant as [t] more often after the back vowels ([u] and [o]) than the front vowels ([i] and [e]). That is, the listeners were likely to identify an ambiguous stop consonant as a segment with the relatively higher spectral frequencies, which is [t], after a segment with the relatively lower formant frequencies, which is either [u] or [o]. Thus, Kingston et al. obtained the basic patterns of perceptual contextual dissimilation for the listeners' broad phonological categories of front and back vowels.

However, within the back vowel category, although the mid back vowel [o] had lower formant frequencies than the vowel [u], English listeners identified an ambiguous stop consonant as [t] more often after [u] than after [o]. If the listeners showed perceptual contextual dissimilation according to the phonetic details of the vowels they heard like the listeners in Whalen's (1989) study did, the listeners should have identified an ambiguous stop consonant as [t] more often after the vowel [o] than the after the vowel [u]. For this, Kingston et al. explained that since the vowels were synthesized based on formant frequencies of Japanese vowels, "[t]he poorer fit to English listeners’ expectations might have forced them to rely instead on their phonological analysis of these vowels" (Kingston et al., 2000, p. 520). That is, Kingston et al. assumed that English listeners assimilated (e.g., Best, 1993, 1995) Japanese vowel [u] to their English vowel /u/, which probably has lower F2 and F3 than those they actually heard, and identified an ambiguous stop consonant according to the English /u/, in other words, according to the phonological analysis of the English /u/.

In sum, in Whalen's study, the listeners showed perceptual contextual dissimilation according to the phonetic details of the segments they heard even though the interacting segments (both target and context segments) were poor fit to the listeners. That is, the listeners heard an ambiguous target segment as having low spectral frequencies when it was followed by a vowel with high formant frequencies, and the listeners heard an ambiguous target segment as having high spectral frequencies when it was followed by a vowel with low formant frequencies. However, in Kingston et al.'s study, the listeners' response pattern was different from what could be expected from Whalen's study. The listeners heard an ambiguous target segment as having low spectral frequencies when the vowel [u] preceded, and the listeners heard an ambiguous target segment as having high spectral frequencies when the vowel [o] preceded. However, the vowel [o] had the lower formant frequencies than the vowel [u] in their study. For this, Kingston et al. claimed that when the interacting segments are poor fit, the listeners show perceptual contextual dissimilation according to the phonological analysis. The issue here is that Whalen's claim that listeners show perceptual contextual dissimilation according to the actual phonetic details of the segments they hear cannot explain the English listeners' response patterns in Kingston et al.'s study. Likewise, Kingston et al.'s explanation that listeners show perceptual contextual dissimilation according to the phonological analysis of the segments they hear cannot explain the English listeners' response patterns in Whalen's study.

2.5. A Possible Solution to the Paradox

As described above, Kingston et al. concluded that the reason English listeners responded [t] more often after the vowel [u] than after the vowel [o] was because they showed perceptual dissimilation according to their phonological analysis of the English /u/. However, there is another way to interpret the English listeners' response patterns in Kingston et al.'s study. Since the vowel [u] that the English listeners heard presumably had higher F2 and F3 frequencies than what they would usually hear, the listeners might have assimilated the vowel [u] to the English /u/, but also detected unusual high formant frequencies in the vowel, and attributed the high frequencies (residues after parsing the expected formant frequencies) to the following stop consonant. Since [t] is more likely to be the origin of the high frequencies than [k] is, English listeners might have responded [t] more often after [u] than after [o]. Thus, the proposed assumption here is that the higher number of [t] responses for the back vowels compared to the front vowels were due to perceptual contextual dissimilation; however, the higher number of [t] responses for the vowel [u] compared to the vowel [o] were probably due to perceptual contextual assimilation.

Granting, the "parsing" mechanism originally proposed by Fowler and her colleagues (Fowler, 1984, 1996; Fowler & Brown, 2000; Fowler & Smith, 1986) was that listeners segment along with coarticulatory lines. However, the assumption here is that listeners can use their native language knowledge about vowels (expected formant frequencies) to parse segments. That is, when listeners are presented with a poor fit vowel, [ω] ($\omega = \text{some vowel}$), they assimilate the vowel to their native vowel /ω/, parse the actual acoustic signal with the expected formant frequencies using their native knowledge of the vowel /ω/, and attribute the residues to the neighboring segment. If a poor fit vowel that the listeners hear has higher formant frequencies than expected, listeners think these high spectral frequencies are coming from the neighboring segment, and identify the neighboring segment as having high spectral frequencies. Figure 3 illustrates this. On the other hand, if a poor fit vowel listeners hear has lower formant frequencies than expected, listeners think these low spectral frequencies are coming from the neighboring segment, and identify the neighboring segment as having low spectral frequencies. Figure 4 illustrates this.

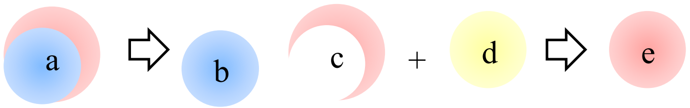

Figure 3. Perceptual contextual assimilation when a vowel has higher formant frequencies than expected
a) the actual phonetic input, a vowel with somewhat higher formant frequencies than the usual one; b) parsed vowel using expected formant frequencies of the vowel from the native knowledge; c) residues after parsing, high spectral frequencies in this case; d) an ambiguous neighboring segment; e) the neighboring segment is identified as having high spectral frequencies.

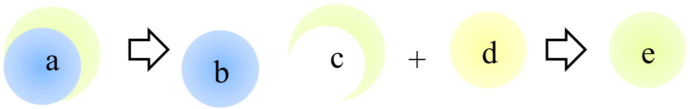

Figure 4. Perceptual contextual assimilation when a vowel has lower formant frequencies than expected
a) the actual phonetic input, a vowel with somewhat lower formant frequencies than the usual one; b) parsed vowel using expected formant frequencies of the vowel from the native knowledge; c) residues after parsing, low spectral frequencies in this case; d) an ambiguous neighboring segment; e) the neighboring segment is identified as having low spectral frequencies.

Although this may be able to explain the response patterns observed in Kingston et al.'s study, this cannot explain the response patterns observed in Whalen's study. Since a poor fit vowel followed an ambiguous consonant in Whalen's study, and a poor fit vowel preceded an ambiguous consonant in Kingston et al.'s study, I hypothesized that listeners might show response patterns like those observed in Whalen for CV syllables, and listeners might show response patterns like those observed in Kingston et al. for VC syllables. This was thought to be plausible based on the claim that anticipatory coarticulation information in a vowel is more easily perceived than carryover coarticulation information in a vowel (Jeong, 2012, however, note that the participants in this study were native Korean speakers). In order to test this assumption against Whalen's Kingston et al.'s claims, an experiment has been carried out.

CHAPTER 3. RESEARCH QUESTIONS, HYPOTHESES, & PREDICTIONS

3.1. Research Questions

The studies discussed above raised two main questions regarding listeners' perceptual response patterns when listeners are presented with CV/VC syllables that are comprised of an ambiguous consonant and a poor fit vowel:

i. Do listeners show perceptual contextual dissimilation or perceptual contextual assimilation for vowels within the same category?

ii. Do listeners show perceptual contextual dissimilation/assimilation according to the phonological analysis of the segments they hear as Kingston et al. (2011) suggested, or according to the actual phonetic details of the segments they hear as Whalen (1989) claimed?

3.2. Hypotheses

As mentioned in the previous chapter, the purpose of the present study was to test a possible solution proposed for the paradoxical results from Whalen's (1989) study and Kingston et al.'s study, as well as Whalen's and Kingston et al.'s claims. Thus, there were 3 hypotheses that were tested:

∗ Hypothesis 1 (based on Kingston et al., 2011): Listeners show perceptual contextual dissimilation according to the phonological analysis of the segments they hear for both the CV and VC stimuli.

- Hypothesis 2 (based on Whalen, 1989): Listeners show perceptual contextual dissimilation according to the actual phonetic details of the segments they hear for both the CV and VC stimuli.
- Hypothesis 3 (an assumption made for the paradox): Listeners show perceptual contextual dissimilation according to the actual phonetic details of the segments they hear for the CV stimuli, but listeners show perceptual contextual assimilation according to the actual phonetic details of the segments they hear for the VC stimuli.

3.3. Experimental Design & Predictions

In order to test these research questions and hypotheses, a *sushi* experiment similar to Experiment 3 of Whalen's study was considered be most appropriate for the present study. Thus, native English listeners' perception of CV/VC syllables comprised of an ambiguous consonant and a poor fit vowel was investigated. The ambiguous consonant was an ambiguous fricative noise drawn from a [s]-[ʃ] continuum, and the poor fit vowel was a vowel drawn from an [i]-[u] continuum following Whalen. Unlike Whalen's stimuli, however, the poor fit vowels were either an [i] that has higher formant frequencies than a good exemplar of English [i], an [i] that has lower formant frequencies than a good exemplar of English [i], an [u] that has higher formant frequencies than a good exemplar of English [u], or an [u] that has lower formant frequencies than a good exemplar of English [u]. In order to test just the effects of the heard order of fricatives and vowels, the VC stimuli were synthesized by literally reversing the CV stimuli following Mann and Soli (1991). The listeners' response choices were 'see,' 'she,' 'sue,' and 'shoe' for the CV stimuli, and 'eece,' 'eesh,' ooce,' and 'oosh' for the VC stimuli.

All of the 3 hypotheses above predict that the listeners would show perceptual contextual dissimilation for the overall picture. That is, all of these hypotheses predict that the listeners would identify an ambiguous fricative noise as [ʃ] more often for the vowel [i] than for the vowel [u] because [ʃ] has its main spectral frequencies in relatively low spectral frequencies and [i] has relatively higher formant frequencies compared to the vowel [u]. These hypotheses, however, predict different response patterns for vowels within the same category as described below.

If Hypothesis 1 is right, listeners should identify an ambiguous fricative noise as [ʃ] more often when the vowel is identified as [i] (thus, /i/) than when the vowel is identified as [u] (thus, /u/) for both the CV and VC stimuli. Moreover, since the listeners are expected to show perceptual contextual dissimilation according to their phonological analysis of the segments they hear, an [i] with higher formant frequencies and an [i] with lower formant frequencies should trigger the same number of [ʃ] responses because both [i]s are the same /i/. Similarly, an [u] with higher formant frequencies and an [u] with lower formant frequencies should trigger the same number of [ʃ] responses because both [u]s are the same /u/. This hypothesis predicts that what matters is the vowel category, and not the actual phonetic details of the segments. The predicted overall response pattern is shown in Figure 5.

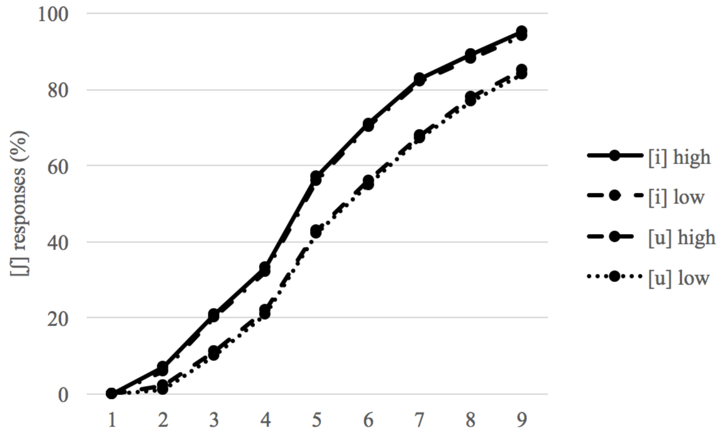

Figure 5. The predicted response pattern of Hypothesis 1 for the CV/VC stimuli
The x-axis represents the consonant continuum, where 1 is the [s]-end and 9 is the [ʃ]-end.

If Hypothesis 2 is right, the listeners should identify an ambiguous fricative noise as [ʃ] more often when the noise is followed/preceded by a vowel with higher formant frequencies than a vowel with lower formant frequencies. That is, the listeners should identify an ambiguous fricative noise as [ʃ] more often when the noise is followed/preceded by an [i] with higher formant frequencies than an [i] with lower formant frequencies. Likewise, the listeners would identify an ambiguous fricative noise as [ʃ] more often when the noise is followed/preceded by an [u] with higher formant frequencies than an [u] with lower formant frequencies. Unlike Hypothesis 1, this hypothesis predicts that what matters is the actual phonetic details of the segments, and not the vowel category. The predicted overall response pattern is shown in Figure 6.

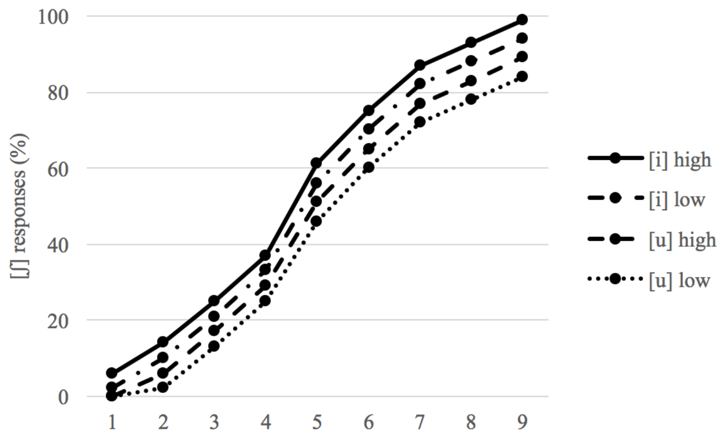

Figure 6. The predicted response pattern of Hypothesis 2 for the CV/VC stimuli
The x-axis represents the consonant continuum, where 1 is the [s]-end and 9 is the [ʃ]-end.

If Hypothesis 3 is right, listeners should identify an ambiguous fricative noise as [ʃ] more often when the noise is followed by a vowel with higher formant frequencies than a vowel with lower formant frequencies for the CV stimuli. This portion is much like what Hypothesis 2 would predict. As for the VC stimuli, this hypothesis still predicts that the listeners would identify an ambiguous fricative noise as [ʃ] more often when the noise is preceded by the vowel [i] than when the noise is preceded by the vowel [u]. However, when vowels within the same category are considered, this hypothesis predicts somewhat opposite of what Hypothesis 2 predicts. For the VC stimuli, this hypothesis predicts that the listeners should identify an ambiguous fricative noise as [ʃ] more often when the noise is preceded by an [i] with lower formant frequencies than an [i] with higher formant frequencies. Likewise, the listeners should identify an ambiguous fricative noise as [ʃ] more often when the noise is preceded by an [u] with lower formant frequencies than an [u] with higher formant frequencies. This is because this hypothesis assumes that when listeners are presented with vowels that have higher formant frequencies than what they would usually hear, listeners would think the high frequencies are due to the following segment, and identify an ambiguous fricative noise as [s]. Likewise, when listeners are presented with vowels that have lower formant frequencies than what they would usually hear, listeners would think the low frequencies are due to the following segment and identify an ambiguous fricative noise as [ʃ]. Figure 7 shows the predicted response pattern for the CV stimuli, and Figure 8 shows the predicted response pattern for the VC stimuli.

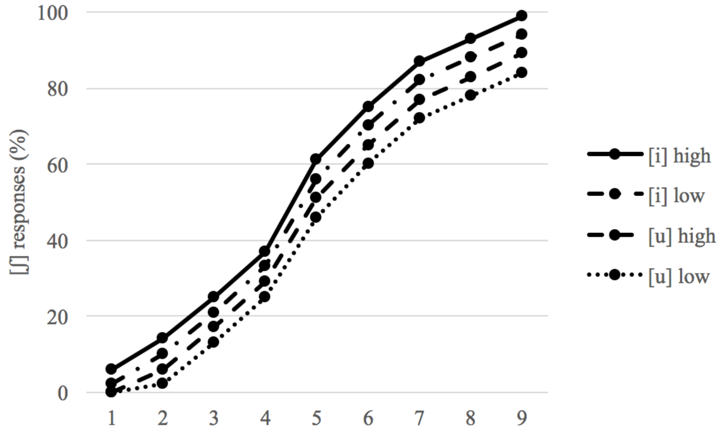

Figure 7. The predicted response pattern of Hypothesis 3 for the CV stimuli
The x-axis represents the consonant continuum, where 1 is the [s]-end and 9 is the [ʃ]-end.

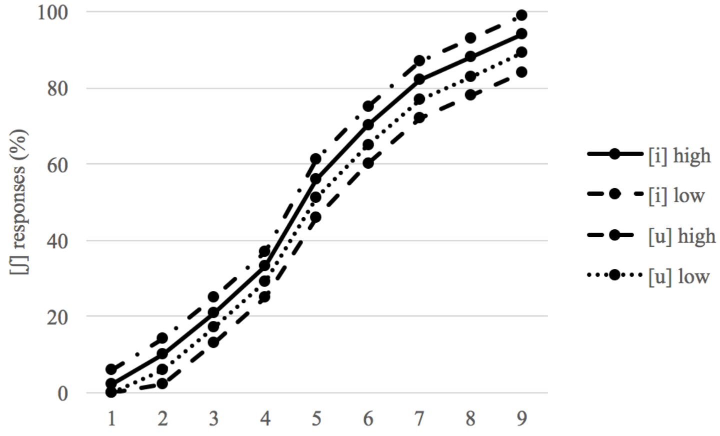

Figure 8. The predicted response pattern of Hypothesis 3 for the VC stimuli
The x-axis represents the consonant continuum, where 1 is the [s]-end and 9 is the [ʃ]-end.

CHAPTER 4. EXPERIMENTAL METHODS

4.1. Stimuli

To test the hypotheses, CV and VC syllables where C is drawn from a [s]-[ʃ] continuum and V is drawn from an [i]-[u] continuum were needed. All the stimuli were synthesized with Praat software 6.0.28. (Boersma & Weenink, 2017).

Recording: A male native speaker of English produced CV and VC syllables where C was either [s] or [ʃ] and V was either [i] or [u] ([si], [ʃi], [su], [ʃu], [is], [iʃ], [us], and [uʃ]) multiple times for each in a quiet office. These utterances were digitized at a sampling rate of 44100 Hz with 16-bit resolution using Praat SoundRecorder. Several of the clearest utterances for each item were chosen for stimuli synthesis.

Fricative synthesis: In order to make both the CV and VC stimuli sound as natural as possible, especially the fricative portion of the VC stimuli when the CV stimuli were reversed, durations of the fricatives of these utterances were measured. The average durations of the fricative noises were 290 msec for CV syllables and 390 msec for VC syllables. Although the durations of fricative noises used in Whalen (1989) were 200 msec before the vowel [u] and 220msec before the vowel [i], these durations were determined to be too short for the CV and VC stimuli in the present study. Thus, the clearest [si] and [ʃi] utterances with fricative noises that had similar, and relatively longer durations (approximately 300 msec) were chosen for the fricative synthesis. From these utterances, exactly 300 msec portions of the fricative noises [s] and [ʃ] were excerpted. The intensity of each noise was modified to 65 dB, and a fade out effect was applied to the last 100 msec. The fade out effect was applied so that the noises would sound more natural in the VC stimuli. Using these noises as the endpoints, a 30-step [s]-[ʃ] was created by blending the two endpoints at different intensity ratios. Since for some reason a 10-step continuum did not produce the noises that can be categorized [s] by native English listeners, a 30-step was chosen to make sure the synthesized continuum consists of fricative noises that sound like [s] and [ʃ] for native English listeners.

Vowel synthesis: All the vowels were synthesized with the Klatt synthesizer (Klatt, 1980) implemented in Praat. In order to synthesize poor fit vowels, formant frequencies of the actual (good) [i]s and [u]s were measured. The average formant frequencies of the actual [i]s were F1 = 355 Hz, F2 = 2452 Hz, F3 = 3033 Hz, and F4 = 3686 Hz. The average formant frequencies of the actual [u]s were F1 = 355 Hz, F2 = 1097, F3 = 2455, and F4 = 3572 Hz. An [i]-[u] continuum that have 2 members with formant frequencies that are similar to those of the actual vowels was needed to be synthesized. The formant frequency values of the [i]-end were set to F1 = 370 Hz, F2 = 3000 Hz, F3 = 3500 Hz, F4 = 4100 Hz, F5 = 4700 Hz, and F6 = 5500 Hz. 6 formants were set for each vowel because the outcome was better than vowels with less formant settings. By decreasing the values of all the formant frequencies except the first formant frequencies, a 26-step [i]-[u] continuum was created. F2 varied in equal steps (130 mels) from 3000 Hz (for the [i]-endpoint) to 643 Hz (for the [u]-endpoint). The mel scale was used so that each vowel has the same psychophysical distance between each other (Grieser & Kuhl, 1989; Iverson & Kuhl, 1995, 2000; Kuhl, 1991). The mels were calculated with Fant's (1968) formula, mel= (1000/log(2))(log(f/1000+1)). F3, F4, F5, and F6 decreased exponentially overall so that each vowel has formant frequencies like those of natural vowels. All the formant frequency values of these vowels are shown in Appendix A. The bandwidths for the [i] end were 300, 300, 300, 200, 200, and 200 Hz for F1, F2, F3, F4, F5, and F6 respectively. The bandwidths for F1 and F2 increased by 10, and the bandwidth for F3 decreased by 5 as the step number increased. Since KlattGrids did not filter out desired frequencies nicely, frequencies between formant frequencies were filtered using a stop Hann band. The smoothing below F1 was 200 Hz, the smoothing between F1 and F2 was 100 Hz, the smoothing between F2 and F3 was 200 Hz, and the smoothing between F3 and F4 was 300 Hz. In addition, F1 was filtered out with a pass Hann band with 300 Hz smoothing for the [i]-end. The smoothing increased by 20 Hz towards the [u]-end. F2 was filtered out with a pass Hann band with 100 Hz smoothing for the [i]-end. The smoothing increased by 10 Hz towards the [u]-end. F3 was filtered out with a pass Hann band with 1000 Hz smoothing for the [i]-end. The smoothing decreased by 38 Hz towards the [u]-end. F4 was filtered out with a pass Hann band with 100 Hz smoothing for the [i]-end. The smoothing decreased by 1 Hz towards the [u]-end. F5 and F6 were also filtered out with a pass Hann band with 300 Hz through out the continuum. These values were determined by checking the spectrum of the actual vowels and the synthesized vowels so that they look alike. The duration of each vowel was set to 300 msec (the same duration as the fricative portion), so that both fricative and vowel portions could sound equally ambiguous. The pitch was 135 Hz at 0 msec and remained the same up to 90 msec. It increased to 145 Hz gradually from 90 msec to 180 msec, and remained the same for the rest.

Selection of fricatives/vowels: The synthesized fricative continuum and vowel continuum were presented to 2 native English listeners to determine the category boundary of [s] and [ʃ] for the fricative continuum and the category boundary of [i] and [u] for the vowel continuum. In addition, these native English listeners were asked to pick an [i] that most sounds like English [i] and an [u] that most sounds like English [u] for the vowel continuum to make sure the vowels that have formant frequency values similar to natural vowels are perceived as most English-like. Based on the actual utterances, the vowels #7 and #8 were supposed to be identified as the good [i], and the vowels #20, and #21 were supposed to be identified as the best [u].

The category boundary of the fricative continuum for one of the listeners was between #11 and #12. The category boundary of the fricative continuum for the other listener was between #9 and #10. Since some pilot participants heard only [s] and some participants heard only [ʃ] during a pilot experiment that used a narrow range of fricative noises from the continuum, it was considered that it is safe to use a wide range of fricative noise from the continuum. Thus, 9 fricative noises, which were #1, 3, 5, 7, 9, 11, 13, 15, and 17 (#1 was the most [s]-like, and #17 was the most [ʃ]-like), were chosen for the CV and VC stimuli. The long term average spectrum (LTAS) of these noises are shown in Figure 9.

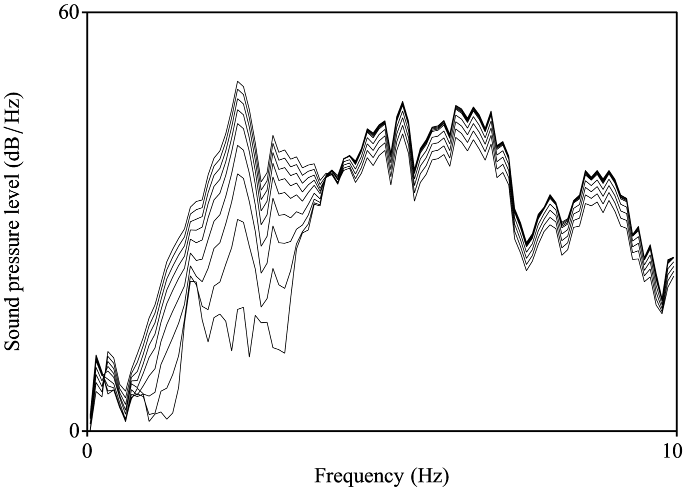

Figure 9. LTAS (long term average spectrum) of the consonant stimuli

As for the vowels, the category boundary of the [i]-[u] continuum for one of the listeners was between #11 and #12. She picked #8 as the best [i] and #20 as the best [u]. The category boundary of the [i]-[u] continuum for the other listener was between #11 and #12. She picked #7 as the best [i] and #21 as the best [u]. The vowels that were needed were as follows: an [i] with higher formant frequencies than a good exemplar of English [i], an [i] with lower formant frequencies than a good exemplar of English [i], an [u] with higher formant frequencies than a good exemplar of English [u], and an [u] with lower formant frequencies than a good exemplar of English [u], where i) an [i] with higher formant frequencies than a good exemplar of English [i] and an [i] with lower formant frequencies than a good exemplar of English [i] share equal distance from a good exemplar of English [i], and an [u] with higher formant frequencies than a good exemplar of English [u] and an [u] with lower formant frequencies than a good exemplar of English [u] share equal distance from a good exemplar of English [u]; and ii) each chosen vowel is an equal number of steps away from one another. Thus, #5 was chosen for an [i] that has higher formant frequencies than a good exemplar of English [i], #11 was chosen for an [i] that has lower formant frequencies than a good exemplar of English [i], #17 was chosen for an [u] that has higher formant frequencies than a good exemplar of English [u], and #23 was chosen for an [u] that has lower formant frequencies than a good exemplar of English [u]. Table 1 shows the formant frequency values of these vowels, as well as a good exemplar of English [i] (#8) and a good exemplar of English [u].

Stimulus #	**F1**	**F2**	**F3**	**F4**	**F5**	**F6**
5	370	2623	3169	3783	4383	5183
8	370	2340	2979	3613	4213	5013
11	370	2057	2829	3486	4086	4886
17	370	1491	2615	3318	3918	4718
20	370	1208	2540	3264	3864	4664
23	370	926	2480	3223	3823	4623

Table 1. The formant frequency values (in Hz) of the vowel stimuli
#5 is an [i] with higher formant frequencies than a good English [i]; #8 is a good English [i], #11 is an [i] with lower formant frequencies than a good English [i]; #17 is an [u] with higher formant frequencies than a good English [u]; #20 is a good English [u], #23 is an [u] with lower formant frequencies than a good English [u].

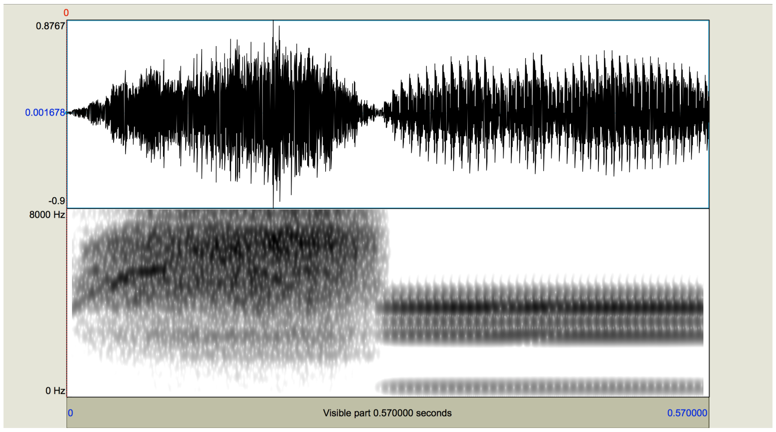

Figure 10. A waveform and spectrogram of an experiment stimulus
The fricative #3 ([s]-like) concatenated with the vowel #5 ([i] with the higher formant frequencies than a good English [i]).

CV/VC stimuli synthesis: Then, each consonant and each vowel were concatenated with a 30 msec overlap, which yielded 36 CV stimuli. By reversing these CV stimuli, 36 VC stimuli were created. Figure 10 shows an example of a stimulus. In addition to these stimuli, using the fricative noises #1 (the most [s]-like stimulus) and #17 (the most [ʃ]-like stimulus), and the vowels #8 and #20 (the good [i] and the good [u]), another set of CV and VC stimuli was synthesized for use in practice tasks.

4.2. Participants

37 participants were recruited from the University of North Carolina at Chapel Hill community, and were paid $8 for participating in an approximately 30-min session. The results of 5 participants were excluded from the data analysis for several reasons (non-native American English speaker, or did not pass the 90% correct response criterion in the practice tasks). All other participants were monolingual, native speakers of American English with no reported hearing impairment. The ages were ranged from 19 years old to 64 years old. There were 22 female and 10 male listeners.

4.3. Procedure

The experiment was done in a sound-proof booth. The experiment was run in the Praat MFC environment on a laptop computer with headphones. There were 2 tasks: Task CV and

Task VC. In Task CV, only CV stimuli were presented. Each CV stimulus was presented 3 times. All the tokens were presented randomly. The participants heard one CV stimulus at a time. The participants were instructed to categorize the stimuli into 'see', 'she', 'sue', or 'shoe' by pressing the corresponding key of a keyboard after hearing each token. In Task VC, only VC stimuli were presented. Each VC stimulus was presented 3 times. All the tokens were presented randomly. The participants heard one VC stimulus at a time. The participants were instructed to categorize the stimuli into 'eece', 'eesh', 'ooce', or 'oosh' by pressing the corresponding key of a keyboard after hearing each token. Each task had a short practice task where only CV stimuli (for Task CV) and VC stimuli (for Task VC) that were synthesized with a good exemplar of English [i] (#8) and a good exemplar of English [u] (#20) were presented. Half of the participants did Task CV first, then Task VC afterwards. The other half of the participants did Task VC first, then Task CV afterwards.

CHAPTER 5. STATISTICAL ANALYSIS & RESULTS

5.1. Preview

In both Task CV and Task VC, both of the vowels #5 (a poor fit [i] with higher formant frequencies than a good English [i]) and #11 (a poor fit [i] with lower formant frequencies than a good English [i]) were heard as [i] most of the times, and both of the vowels #17 (a poor fit [u] with higher formant frequencies than a good English [u]) and #23 (a poor fit [u] with lower formant frequencies than a good English [u]) were heard as [u] most of the times. The fricative noises around the [s]-end were heard as [s] most of the time, and fricative noises around the [ʃ]-end were heard as [ʃ] most of the times. Each stimulus was identified 96 times in total (3 repetitions x 33 participants). Table 2 shows the raw numbers of responses for the CV stimuli when an ambiguous vowel was identified as [i]. Table 3 shows the raw numbers of responses for the CV stimuli when an ambiguous vowel was identified as [u]. Table 4 shows the raw numbers of responses for the VC stimuli when an ambiguous vowel was identified as [i]. Table 5 shows the raw numbers of responses for the VC stimuli when an ambiguous vowel was identified as [u]. Percentages of these responses are shown in Figure 11, 12, 13, and 14 respectively.

	Vowel 5	Vowel 11	Vowel 17	Vowel 23
Fricative 1	0	1	0	0
Fricative 3	4	5	1	0
Fricative 5	22	21	2	0
Fricative 7	58	61	5	1
Fricative 9	70	77	7	1
Fricative 11	79	82	4	1
Fricative 13	80	89	7	0
Fricative 15	85	94	9	1
Fricative 17	89	91	6	0

Table 2. The raw [ʃ] responses for the CV stimuli when a poor fit vowel was identified as [i]

There were 96 responses for each stimulus. Fricative 1 was most [s]-like, and Fricative 17 was most [ʃ]-like. Vowel 5 was the [i] with higher formant frequencies; Vowel 11 was the [i] with lower formant frequencies; Vowel 17 was the [u] with higher formant frequencies; and Vowel 18 was the [u] with lower formant frequencies.

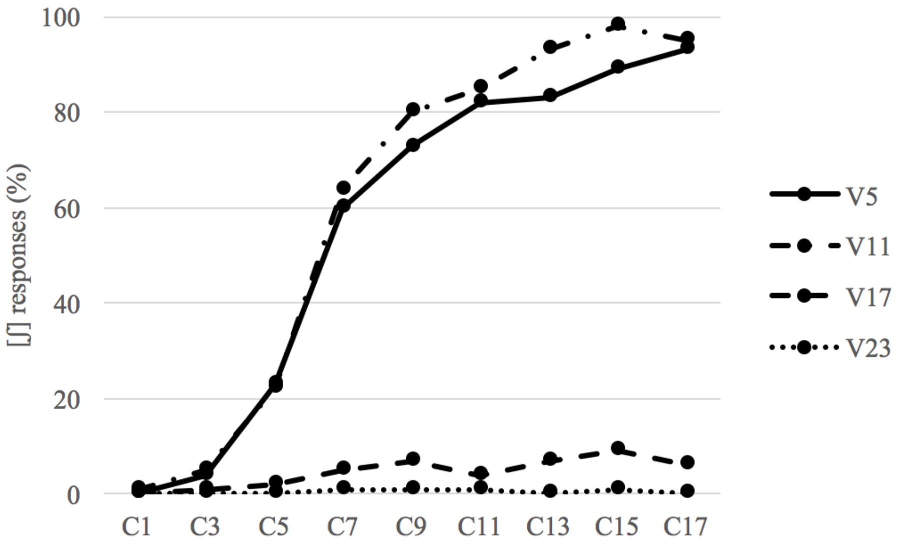

Figure 11. The percentages of [ʃ] responses for the CV stimuli when a poor fit vowel was identified as [i]

Fricative 1 was most [s]-like, and Fricative 17 was most [ʃ]-like. Vowel 5 was the [i] with higher formant frequencies; Vowel 11 was the [i] with lower formant frequencies; Vowel 17 was the [u] with higher formant frequencies; and Vowel 18 was the [u] with lower formant frequencies.

	Vowel 5	Vowel 11	Vowel 17	Vowel 23
Fricative 1	0	0	3	13
Fricative 3	0	0	2	19
Fricative 5	0	0	16	41
Fricative 7	0	0	30	57
Fricative 9	0	0	40	68
Fricative 11	0	0	58	69
Fricative 13	2	1	56	77
Fricative 15	3	1	58	81
Fricative 17	2	0	70	84

Table 3. The raw [ʃ] responses for the CV stimuli when a poor fit vowel was identified as [u]

There were 96 responses for each stimulus. Fricative 1 was most [s]-like, and Fricative 17 was most [ʃ]-like. Vowel 5 was the [i] with higher formant frequencies; Vowel 11 was the [i] with lower formant frequencies; Vowel 17 was the [u] with higher formant frequencies; and Vowel 18 was the [u] with lower formant frequencies.

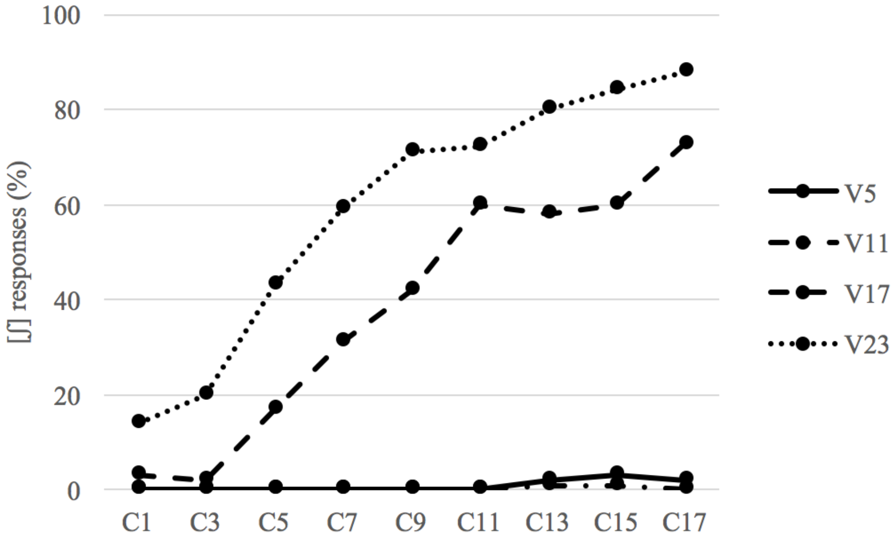

Figure 12. The percentages of [ʃ] responses for the CV stimuli when a poor fit vowel was identified as [u]

Fricative 1 was most [s]-like, and Fricative 17 was most [ʃ]-like. Vowel 5 was the [i] with higher formant frequencies; Vowel 11 was the [i] with lower formant frequencies; Vowel 17 was the [u] with higher formant frequencies; and Vowel 18 was the [u] with lower formant frequencies.

	Vowel 5	Vowel 11	Vowel 17	Vowel 23
Fricative 1	0	0	1	0
Fricative 3	6	8	0	0
Fricative 5	34	50	2	0
Fricative 7	56	66	3	0
Fricative 9	72	84	5	0
Fricative 11	85	81	6	0
Fricative 13	87	87	4	0
Fricative 15	91	86	4	2
Fricative 17	92	89	2	0

Table 4. The raw [ʃ] responses for the VC stimuli when a poor fit vowel was identified as [i]

There were 96 responses for each stimulus. Fricative 1 was most [s]-like, and Fricative 17 was most [ʃ]-like. Vowel 5 was the [i] with higher formant frequencies; Vowel 11 was the [i] with lower formant frequencies; Vowel 17 was the [u] with higher formant frequencies; and Vowel 18 was the [u] with lower formant frequencies.

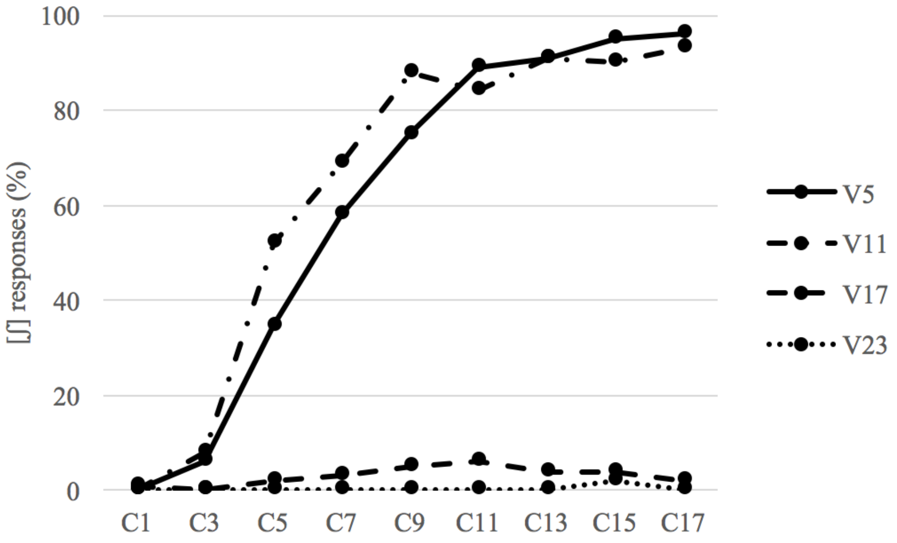

Figure 13. The percentages of [ʃ] responses for the VC stimuli when a poor fit vowel was identified as [i]

Fricative 1 was most [s]-like, and Fricative 17 was most [ʃ]-like. Vowel 5 was the [i] with higher formant frequencies; Vowel 11 was the [i] with lower formant frequencies; Vowel 17 was the [u] with higher formant frequencies; and Vowel 18 was the [u] with lower formant frequencies.

	Vowel 5	Vowel 11	Vowel 17	Vowel 23
Fricative 1	0	0	1	2
Fricative 3	0	0	10	18
Fricative 5	0	1	44	48
Fricative 7	0	3	51	71
Fricative 9	0	3	74	84
Fricative 11	0	4	80	86
Fricative 13	1	1	87	93
Fricative 15	0	4	83	93
Fricative 17	0	2	92	94

Table 5. The raw [ʃ] responses for the VC stimuli when a poor fit vowel was identified as [u]

There were 96 responses for each stimulus. Fricative 1 was most [s]-like, and Fricative 17 was most [ʃ]-like. Vowel 5 was the [i] with higher formant frequencies; Vowel 11 was the [i] with lower formant frequencies; Vowel 17 was the [u] with higher formant frequencies; and Vowel 18 was the [u] with lower formant frequencies.

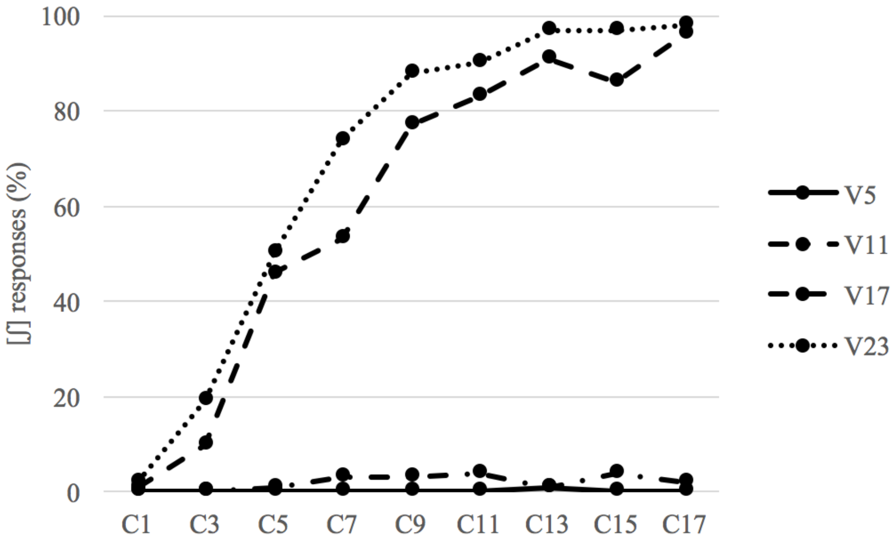

Figure 14. The percentages of [ʃ] responses for the VC stimuli when a poor fit vowel was identified as [u]

Fricative 1 was most [s]-like, and Fricative 17 was most [ʃ]-like. Vowel 5 was the [i] with higher formant frequencies; Vowel 11 was the [i] with lower formant frequencies; Vowel 17 was the [u] with higher formant frequencies; and Vowel 18 was the [u] with lower formant frequencies.

5.2. Statistical Analysis & Results

The listeners' responses were analyzed using mixed-effects logistic regression with [ʃ] responses as the dependent variable ([ʃ] = 1, [s] = -1). The model includes subjects as a random factor. The fixed factors included in the model are as follows: FFreq (continuous), which refers to 9 fricative stimuli, in other words, fricative spectral frequencies; VCat (categorical), which refers to the vowel category responses from the participants, [u] or [i]; VFreq (continuous), which refers to 4 vowel stimuli, in other words, vowel formant frequencies; CV/VC (categorical), which refers to the order of the consonant and the vowel, CV syllables or VC syllables. The coded values used for these fixed effects were as follows: -4, -3, -2, -1, 0, 1, 2, 3, and 4 for FFreq (more [s]-like = -4, more [ʃ]-like = 4); 1 and -1 for VCat ([u] = 1, [i] = -1); -3, -1, 1, and 3 for VFreq (more [i]-like = -3, more [u]-like = 3); and 1 and -1 for CV/VC (CV = 1, VC = -1). The following interaction terms were also included in the model: VFreq:VCat, VCat:CV/VC, VFreq:CV/VC, FFreq:CV/VC, and VFreq:VCat: CV/VC. The results of regression are shown in Table 6.

Parameter	**Estimate**	**Standard error**	**95% lower confidence limit**	**95% upper confidence limit**	**z-value**	**p-value**
Intercept	0.4821	0.2124	0.0658	0.8984	2.27	0.0232
FFreq	0.7959	0.0702	0.6582	0.9335	11.33	<.0001
VCat	-0.5374	0.1065	-0.7463	-0.3286	-5.04	<.0001
VFreq	0.2347	0.0440	0.1486	0.3209	5.34	<.0001
CV/VC	-0.6966	0.1479	-0.9865	-0.4068	-4.71	<.0001
VFreq:CV/VC	0.0291	0.0535	-0.0757	0.1339	0.54	0.5859
VCat:CV/VC	-0.2459	0.1187	-0.4786	-0.0133	-2.07	0.0383
FFreq:CV/VC	-0.1200	0.0581	-0.2339	-0.0062	-2.07	0.0387
VFreq:VCat	0.1586	0.0372	0.0858	0.2314	4.27	<.0001
VFreq:VCat:CV/VC	0.1214	0.0288	0.0650	0.1779	4.21	<.0001

Table 6. Results of mixed-effects logistic regression

The positive intercept reveals overall bias towards [ʃ] responses. As FFreq increased towards 3 (as fricative spectral frequencies became lower), the number of [ʃ] responses increased significantly ($z = 11.33$, $p < .0001$). The negative estimate value of VCat reveals there was a significant association between [ʃ] and [i], in other words, [s] and [u] ($z = -5.04$, $p < .0001$). The positive estimate value of VFreq reveals as VFreq increased towards 3 (as vowel formant frequencies became lower), [ʃ] responses increased significantly ($z = 5.34$, $p < .0001$). The negative estimate value of CV/VC reveals there was a significant association between [ʃ] responses and VC syllables, in other words [s] responses and CV syllables ($z = -4.71$, $p < .0001$). The interaction effects of VFreq and CV/VC, VCat and CV/VC, and FFreq and CV/VC were not significant ($z = 0.54$, $p = 0.5859$ for the VFreq and CV/VC interaction, $z = -2.07$, $p = 0.0383$ for the VCat and CV/VC interaction, and $z = -2.07$, $p = 0.0387$ for the FFreq and CV/VC interaction). However, the interaction effect of VFreq and VCat was significant ($z = 4.27$, $p < .0001$). The interaction effect of VFreq, VCat, and CV/VC was significant ($z = 4.21$, $p < .0001$).

In further analysis, a comparison between VCat = 1 (when an ambiguous vowel was identifies as [u]) and VCat = -1 (when an ambiguous vowel was identified as [i]) revealed that, overall (for both the CV and VC stimuli), the listeners heard an ambiguous fricative noise as [ʃ] more often when it was followed/preceded by an ambiguous vowel that was categorized as [i] than an ambiguous vowel that was categorized as [u]. The listeners identified an ambiguous fricative noise as [ʃ] 73.5% of the time when an ambiguous vowel was identified as [i], but the listeners identified an ambiguous fricative noise as [ʃ] 48.6% of the time when an ambiguous vowel was identified as [u]. This difference was significant ($\chi 2 = 25.44$, $p < .0001$). For the CV stimuli only, the listeners heard an ambiguous fricative noise as [ʃ] 63.9% of the time when an ambiguous vowel was identified as [i], and the listeners identified an ambiguous fricative noise as [ʃ] 26.9% of the time when an ambiguous vowel was identified as [u]. This difference was significant ($\chi2 = 17.41$, $p < .0001$). For the VC stimuli, the listeners identified an ambiguous fricative noise as [ʃ] 81.3% when an ambiguous vowel was identified as [i], and the listeners identified an ambiguous fricative noise as [ʃ] 70.8% when an ambiguous vowel was identified as [u]. This difference was significant ($\chi2 = 5.44$, $p = 0.0197$). Thus, the effect of the vowel category [i] triggering [ʃ] responses was larger for the CV stimuli than the VC stimuli.

The comparison between VFreq with the smaller value and VFreq with the larger value revealed that the listeners identified an ambiguous fricative noise as [ʃ] more often for the vowels with the lower formant frequencies than for the vowels with the higher formant frequencies ($\chi2 = 28.52$, $p < .0001$). This tendency was larger for the CV stimuli than the VC stimuli ($\chi2 = 9.90$, $p = 0.0017$ for the CV stimuli, $\chi2 = 16.60$, $p < 0.001$ for the VC stimuli). This was due to the effect of VFreq -3 triggering more [ʃ] responses than VFreq -1 and the effect of VFreq 3 triggering more [ʃ] responses than VFreq 1.

After a comparison between VFreq with the smaller value and VFreq with the larger value when an ambiguous vowel was identified as [i] (VCat = -1), it was revealed that the listeners were equally likely to identify an ambiguous fricative noise as [ʃ] for vowels with lower formant frequencies and for vowels with higher formant frequencies when these were identified as [i] for the CV stimuli as difference between [ʃ] responses for vowels with lower formant frequencies and for vowels with higher formant frequencies was not significant ($\chi2 = 1.71$, $p = 0.1916$). However, the listeners identified an ambiguous fricative noise as [ʃ] more often for vowels with lower formant frequencies than for vowels with higher formant frequencies when they were identified as [i] for the VC stimuli ($\chi2 = 6.83$, $p = 0.0090$).

A comparison between VFreq with the smaller value and VFreq with the larger value when an ambiguous vowel was identified as [u] (VCat = 1) revealed that the listeners identified an ambiguous fricative noise as [ʃ] more often for vowels with lower formant frequencies than for vowels with higher formant frequencies when they were identified as [u] for the CV stimuli ($\chi 2 = 27.66$, $p < .0001$). The listeners also identified an ambiguous fricative noise as [ʃ] more often for vowels with lower formant frequencies than for vowels with higher formant frequencies when an ambiguous vowel was identified as [u] for the VC stimuli ($\chi 2 = 15.21$, $p < .0001$).

5.3. Interpretation

The statistical analysis above was interpreted as follows. Overall, the listeners identified an ambiguous fricative noise as [ʃ] more often after and before the vowel category [i] than the vowel category [u]. That is, the listeners were likely to identify an ambiguous segment as a sound with relatively lower frequencies (in this case, [ʃ]) after/before a sound with relatively higher frequencies (in this case, [i]). Thus, the listeners have shown perceptual contextual dissimilation for their broad (more abstract) phonological categories of [i] and [u].

The difference between [ʃ] responses for vowels with lower formant frequencies and for vowels with higher formant frequencies was not significant when a poor fit vowel was identified as [i] for the CV stimuli. Thus, the listeners were not sensitive to phonetic difference between an [i] with higher formant frequencies and an [i] with lower formant frequencies for the CV stimuli.

On the other hand, the difference between [ʃ] responses for vowels with lower formant frequencies and for vowels with higher formant frequencies were significant when a poor fit vowel was identified as [i] for the VC stimuli, and when a poor fit vowel was identified as [u] for the CV and VC stimuli. For these stimuli, the listeners identified an ambiguous fricative noise as [ʃ] more often for the vowels with lower formant frequencies than for the vowels with higher formant frequencies when vowels within the same category are concerned. Thus, the listeners showed perceptual contextual assimilation according to the phonetic details when a poor fit vowel was identified as [i] for the VC stimuli. The listeners also showed perceptual contextual assimilation according to the phonetic details when a poor fit vowel was identified as [u] for both the CV and VC stimuli. The table below summarizes this interpretation.

	CV stimuli	**VC stimuli**
[i]	Not sensitive to phonetic details	Perceptual contextual assimilation according to the phonetic details
[u]	Perceptual contextual assimilation according to the phonetic details	Perceptual contextual assimilation according to the phonetic details

Table 7. Summary of interpretation of the statistical analysis

CHAPTER 6. DISCUSSION

6.1. General Discussion

The basic pattern of perceptual contextual dissimilation was obtained in this study. Overall, the listeners identified an ambiguous fricative noise as [ʃ] more often after and before the vowel category [i] than the vowel category [u]. In addition, the effect of perceptual contextual dissimilation was larger for the CV stimuli than for the VC stimuli. This pattern was consistent with what was reported by past studies (Kunisaki & Fujisaki, 1977; Mann & Soli, 1991).

Mann and Soli (1991) investigated the effects of formant transitions in CV and VC syllables where C was an ambiguous fricative noise from a [s]-[ʃ] continuum and V was an unambiguous vowel with transitions. In the first experiment, they found that the effects of vowels were greater for CV syllables than for VC syllables. When CV and VC syllables were reversed and presented to the listeners in the later experiment, the results were that, again, vocalic effects were greater for CV syllables than for VC syllables. From this, Mann and Soli concluded that later vowels play a greater role. The most plausible explanation for this asymmetrical vocalic contextual effects that Mann and Soli provided was that phonological rules that assimilate a segment to its following segment are much more common for CV syllables than VC syllables (Javkin, 1977). However, Mann and Soli's (1991) acoustic analysis of CV and VC utterances showed no significant difference between these syllables that could explain this perceptual asymmetry. An additional investigation that directly investigates this asymmetrical vocalic contextual effect is needed to reach a more solid conclusion.

Although this study replicated the basic pattern of perceptual contextual dissimilation, none of the hypotheses tested alone can fully explain the results. Hypothesis 1 stated that the listeners would show perceptual contextual dissimilation according to the phonological analysis of the segments they heard for both the CV and VC stimuli. That is, the listeners should be insensitive to the phonetic details of the segments they hear. This hypothesis predicted that listeners should identify an ambiguous fricative noise as [ʃ] more often when the vowel is identified as [i] (thus, /i/) than when the vowel is identified as [u] (thus, /u/) for both the CV and VC stimuli. Although the listeners did not show perceptual contextual dissimilation according to the phonological analysis of the segments they heard as predicted, the listeners have shown perceptual contextual dissimilation for the phonological vowel categories [i] and [u]. However, the listeners were mostly sensitive to the phonetic details, except for when a poor fit vowel was identified as [i] for the CV stimuli.

Hypothesis 2 stated that the listeners would show perceptual contextual dissimilation according to the phonetic details of the segments they hear. This hypothesis predicted that the listeners should identify an ambiguous fricative noise as [ʃ] more often when the noise is followed/preceded by a vowel that has higher formant frequencies than a vowel that has lower formant frequencies. Most of the results, however, showed the opposite pattern. In the present study, the [i] with lower formant frequencies triggered more [ʃ] responses than the [i] with higher formant frequencies, and the [u] with lower formant frequencies triggered more [ʃ] responses than the [u] with higher formant frequencies.

The fact that listeners showed completely no perceptual contextual dissimilation according to the phonetic details of the segments as Whalen (1989) observed was a somewhat surprising result. This might be due to a difference between the vowel stimuli used in Whalen's (1989) study, and the vowel stimuli used in the present study. The vowels used in Whalen's (1989) study were quite ambiguous, as they were around the category boundary between [i] and [u]. For example, an ambiguous vowel used in Whalen (1989) was heard as [i] 56.5% of the time and as [u] 43.5% of the time in one of the stimuli. On the other hand, the vowels used in the present study were poor fit, but not ambiguous. Indeed, poor [i]s were heard as [i] most of the time and poor [u]s were heard as [u] most of the time in the present study. It might be that listeners show perceptual contextual dissimilation according to the phonetic details of the segments when these interacting segments are very ambiguous and do not fall into any categories easily.

Hypothesis 3 stated that the listeners would show perceptual contextual dissimilation according to the phonetic details of the segments they hear for the CV stimuli, and the listeners would show perceptual contextual assimilation according to the phonetic details of the segments they hear for the VC stimuli. This hypothesis predicted that the listeners should identify an ambiguous fricative noise as [ʃ] more often when the noise is followed by a vowel with higher formant frequencies than a vowel with lower formant frequencies for the CV stimuli, but the listeners should identify an ambiguous fricative noise as [ʃ] more often when the noise is preceded by a vowel with lower formant frequencies than a vowel with higher formant frequencies for the VC stimuli. The assumption for the VC stimuli was correct. As predicted, the results show that an [i] with lower formant frequencies triggered more [ʃ] responses than an [i] with higher formant frequencies, and an [u] with lower formant frequencies triggered more [ʃ] responses than an [u] with higher formant frequencies for the VC stimuli. However, the listeners also showed perceptual contextual assimilation for the CV stimuli when a poor fit vowel was identified as [u]. That is, an [u] with lower formant frequencies also triggered more [ʃ] responses than an [u] with higher formant frequencies for the CV stimuli. Moreover, the listeners did not show perceptual contextual dissimilation according to the phonetic details of the segments they heard for the CV stimuli when a poor fit vowel was identified as [i].

The results that the listeners identified an ambiguous fricative noise as [ʃ] more often for an [i] with lower formant frequencies than a good English [i] and [s] more often for an [i] with higher formant frequencies than a good English [i] for the VC stimuli imply that the listeners presumably assimilated the vowel [i] they heard to their native vowel /i/, parsed the actual acoustic signal with the expected formant frequencies using their native knowledge of the vowel /i/, and attributed the residues to the neighboring segment. Likewise, the results that the listeners identified an ambiguous fricative noise as [ʃ] more often for an [u] with lower formant frequencies than a good English [u] and [s] more often for an [u] with higher formant frequencies than a good English [i] for both CV and VC stimuli imply that the listeners presumably assimilated the vowel [u] they heard to their native vowel /u/, parsed the actual acoustic signal with the expected formant frequencies using their native knowledge of the vowel /u/, and attributed the residues to the neighboring segment.

The results that the listeners were insensitive to the acoustic difference between the [i] with higher formant frequencies and the [i] with lower formant frequencies for the CV stimuli are unexpected. This is difficult to interpret for several reasons: First, it is difficult to interpret because the listeners were insensitive to phonetic details only when a poor vowel was identified as [i] and only for the CV stimuli. Second, it is difficult to interpret because, although the listeners were insensitive to phonetic details for the CV stimuli when a poor vowel was identified as [i], the listeners showed the stronger effects of perceptual contextual assimilation according to the phonetic details of the segments they heard when a poor vowel was identified as [u] for the CV stimuli. That is, the listeners were more sensitive to the phonetic details of the segments they heard for the CV stimuli than the VC stimuli when a poor fit vowel was identified as [u].

The stimuli that the listeners showed this response pattern to were the English words 'see' ([si]) and 'she' ([ʃi]), which both occur more frequently than the English words, 'sue' ([su]) and 'shoe' ([ʃu]), and the VC syllables ([is], [iʃ], [us], and [uʃ]). Listeners might have a tendency to be insensitive to the phonetic details of the segments they hear for more familiar words than for less familiar words. In other words, listeners might hear a frequent word as a whole, and not using "the phonetic mode of listening" (Johnson, 2002). However, in the recent study done by White et al. (2013), the listeners' familiarity increases sensitivity to phonetic details if differences in phonetic details are large enough. Since the listeners in the present study have shown perceptual contextual assimilation according to the phonetic details for VC syllables when an ambiguous vowel was categorized as [i], the difference between an [i] with higher formant frequencies and an [i] with lower formant frequencies should have been large enough. Then, the listeners' sensitivity to phonetic details could have been increased for the English words 'see' ([si]) and 'she' ([ʃi]) according to White et al. (2013).

6.2. Some Theoretical Implications

These results imply some important theoretical implications. First, recent studies on perceptual contextual dissimilation often utilize foreign languages. If foreign vowels are used in these studies, however, the results in the present study indicate that listeners might show both perceptual contextual dissimilation and perceptual contextual assimilation. In other words, the results obtained in these perceptual contextual dissimilation studies may not reflect pure perceptual contextual dissimilation effects, because the perception of consonants may be affected by perceptual contextual assimilation due to non-native vowels.

This may be true for second language studies. For example, Takagi and Mann (1955) investigated Japanese listeners' perception of English liquids [l] and [ɹ]. In their study, the listeners were likely to hear the liquid [l] as the liquid [ɹ] especially before the English vowels [u] and [ɑ]. This bias might be because the Japanese listeners assimilated these vowels to Japanese vowels, /ɯ/ and /a/ (which have the higher formant frequencies than English vowels, [u] and [ɑ]), and thus, attributed the low formant frequencies in the English vowels [u] and [ɑ] to the preceding liquid, which may have caused the liquid [l] to sound more [ɹ]-like to the Japanese listeners.

The present results also imply that the definition of "parsing" might need to be changed. As mentioned in Chapter 2, Fowler and her colleagues (Fowler, 1984, 1996; Fowler & Brown, 2000; Fowler & Smith, 1986) claimed that listeners parse acoustic information along with "gestural lines." However, The results of the present study indicate that listeners might be able to parse vowels using their native knowledge.

If this contextual perceptual assimilation mechanism on vowels is acting as a repair mechanism for distorted segments, it would be significant to the theory of sound-change developed by Ohala (1981, 2012). In his theory, sound change is due to "hypocorrection" (listeners' failure of correcting coarticulatory effects) or "hypercorrection" (listeners' unnecessary application of correction) of ambiguous segments. The "correction," in Ohala's theory is based on the perceptual contextual dissimilation mechanism. For example, Ohala (1981, 2012) claims that the utterance of [yt] may be wrongly considered as a distorted form of /ut/, and listeners may accept the utterance [yt] as /ut/ using perceptual contextual dissimilation, which eventually causes a sound change of the vowel [y] to the vowel [u]. However, the results in the present study indicate that listeners are likely to show perceptual contextual assimilation rather than dissimilation in this kind of scenario.

However, the results of the present study indicate that listeners might show perceptual contextual dissimilation according to the phonetic details only if vowels are truly ambiguous, and cannot be categorized easily. The distorted vowels in Ohala's (1981, 2012) scenario, however, are likely to be poor fit rather than ambiguous. If the utterance of [yt] may be wrongly considered as it is a distorted form of /ut/, and it is wrongly corrected by listeners, listeners are likely to do so using perceptual contextual assimilation rather than perceptual contextual dissimilation. That is, the utterance of [yt] may be wrongly corrected to [up] by parsing the vowel [y] with the vowel /u/, and attribute extra high spectral frequencies to the following consonant, which makes the stop consonant [t] sounds more like the stop consonant [p].

6.3. Conclusion

In the present study, native English listeners' perception of an ambiguous fricative noise from a [s]-[ʃ] continuum followed/preceded by poor fit vowels was investigated. The poor fit vowels were an [i] with higher formant frequencies than a good English [i], an [i] with lower formant frequencies than a good English [i], an [u] with higher formant frequencies than a good English [u], and an [u] with lower formant frequencies than a good English [u]. There were two main research questions that were addressed. First, the present study investigated if listeners show perceptual contextual dissimilation or perceptual contextual assimilation. Second, the present study investigated if listeners show perceptual contextual dissimilation/assimilation according to the phonological analysis of the segments they hear, or according to the phonetic details of the segments they hear.

The results were that the listeners showed perceptual contextual dissimilation for their broad (more abstract) phonological categories [i] and [u]. The listeners identified an ambiguous fricative noise as [ʃ] more often when the noise was followed/preceded by the vowel [i] than the vowel [u]. However, the statistical analysis has shown that the listeners also identified an ambiguous fricative noise as [ʃ] for the vowels with the lower formant frequencies. This was because an [i] with lower formant frequencies than a good English [i] triggered more [ʃ] responses than an [i] with higher formant frequencies than a good English [i] for the VC stimuli, and an [u] with lower formant frequencies than a good English [u] triggered more [ʃ] responses than an [u] with higher formant frequencies than a good English [u] for the CV stimuli. That is, the listeners showed perceptual contextual assimilation according to the phonetic details of the segments they heard for these stimuli. These results indicated that listeners may be able to parse vowels using the native language knowledge and dynamically adjust the acoustic discrepancy by showing perceptual contextual assimilation.

There are some remaining questions, however. The present study replicated the basic effects of perceptual contextual dissimilation. In addition, the effects were larger for the CV stimuli than for the VC stimuli, which was also consistent with past studies (Kunisaki & Fujisaki, 1977; Mann & Soli, 1991). Why this asymmetry happens, however, has yet to be answered. This, as well as the mechanism of perceptual contextual dissimilation, should be further investigated.

Also, why the listeners were not sensitive to acoustic difference between an [i] with higher formant frequencies than a good English [i] and an [i] with lower formant frequencies than a good English [i] for the stimuli identified as [si] and [ʃi] remains unanswered. The discussion above briefly proposed that it may have something to do with word frequency. The relation of listeners' attention of phonetic details and word frequency is also worth investigating in the future.

APPENDIX A

Number	F1	F2	F3	F4	F5	F6
1	370	3000	3500	4100	4700	5500
2	370	2906	3407	4009	4609	5409
3	370	2811	3322	3926	4526	5326
4	370	2717	3242	3851	4451	5251
5	370	2623	3169	3783	4383	5183
6	370	2529	3101	3721	4321	5121
7	370	2434	3038	3664	4264	5064
8	370	2340	2979	3613	4213	5013
9	370	2246	2925	3567	4167	4967
10	370	2151	2875	3524	4124	4924
11	370	2057	2829	3486	4086	4886
12	370	1963	2786	3450	4050	4850
13	370	1868	2746	3419	4019	4819
14	370	1774	2710	3390	3990	4790
15	370	1680	2676	3363	3963	4763
16	370	1586	2644	3339	3939	4739
17	370	1491	2615	3318	3918	4718
18	370	1397	2588	3298	3898	4698
19	370	1303	2563	3280	3880	4680
20	370	1208	2540	3264	3864	4664
21	370	1114	2518	3249	3849	4649
22	370	1020	2498	3235	3835	4635
23	370	926	2480	3223	3823	4623
24	370	831	2463	3212	3812	4612
25	370	737	2447	3202	3802	4602
26	370	643	2433	3192	3792	4592

Table 8. Formant frequency values (in Hz) of all the synthesized vowels
#1 is the [i]-end, #26 is the [u]-end

REFERENCES

Beddor, P. S., McGowan, K. B., Boland, J. E., Coetzee, A. W., & Brasher, A. (2013). The time course of perception of coarticulation. *The Journal of the Acoustical Society of America, 133(4)*, 2350-2366.

Bell-Berti, F. & Harris, K. S. (1979). Anticipatory coarticulation: Some implications from a study of lip rounding. *The Journal of the Acoustical Society of America, 65*, 1268-1270.

Best, C. T. (1993). Emergence of language-specific constraints in perception of non-native speech: A window on early phonological development. In B. de Boysson-Bardies, S. de Schonen, P. Jusczyk, P. MacNeilage, & J. Morton (eds.), *Developmental neurocognition: Speech and face processing in the first year of life* (pp. 289-304). Dordrecht, The Netherlands: Academic Publishers B. V.

Best, C. T. (1995). A direct realist perspective on cross-language speech perception. In W. Strange & J. J. Jenkins (eds.), *Cross-language speech perception* (pp. 171-204). Timonium, MD: York Press.

Boersma, P. & Weenink, D. (2017). *Praat: doing phonetics by computer* [Computer program]. Version 6.0.28, retrieved 28 March 2017 from http://www.praat.org/

Dahan, D., Magnuson, J. S., Tanenhaus, M. K., & Hogan, E. M. (2001). Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition. *Language and Cognitive Processes, 16*, 507–534.

Elman, J. L. & McClelland, J. L. (1986). Exploiting lawful variability in the speech. In J. S. Perkell, & D. H. Klatt, (eds.), *Invariance and variability of speech processes* (pp. 360-385). Hillsdale, NJ: Lawrence Erlbaum Associates.

Elman, J. L. & McClelland, J. L. (1988). Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. *Journal of Memory and Language, 27*, 143-165.

Fant, G. (1968). Analysis and synthesis of speech processes. In B. Malmberg (ed.), *Manual of Phonetics* (pp. 173-276). Amsterdam, North-Holland Publishing Company.

Fowler, C. A. (1984). Segmentation of coarticulated speech in perception. *Perception & Psychophysics, 36*, 359–368.

Fowler, C. A. (1996). Listeners do hear sounds, not tongues. *The Journal of the Acoustical Society of America, 99*, 1730–1741.

Fowler, C. A., & Brown, J. B. (2000). Perceptual parsing of acoustic consequences of velum lowering from information for vowels. *Perception & Psychophysics, 62*, 21-32.

Fowler, C., & Smith, M. (1986). Speech perception as "vector analysis": An approach to the problems of segmentation and invariance. In J. Perkell & D. Klatt (eds.), *Invariance and variability of speech processes* (pp. 123-136). Hillsdale, NJ: Erlbaum.

Fowler, C.A. (2006). Compensation for coarticulation reflects gesture perception, not spectral contrast. *Perception & Psychophysics, 68(2)*, 161–177.

Grieser, D. & Kuhl, P. K. (1989). Categorization of speech by infants: Support for speech-sound prototypes. *Developmental Psychology, 25(4)*, 577-588.

Holt, L. L. & Lotto, A. (2002). Behavioral examinations of the level of auditory processing of speech context effects. *Hearing Research, 167*, 156–169.

Iverson, P. & Kuhl, P. K. (1995). Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. *The Journal of the Acoustical Society of America, 97*, 553–562.

Iverson, P. & Kuhl, P. K. (2000). Perceptual magnet and phoneme boundary effects in speech perception: do they arise from a common mechanism? *Perception & Psychophysics, 62*, 874–886.

Javkin, H. R. (1979). Phonetic universals and phonological change. Ph.D. Dissertation, University of California, Berkeley.

Jeong, S. (2012). Directional asymmetry in nasalization: a perceptual account. *Studies in Phonetics, Phonology and Morphology, 18*, 437-469.

Johnson, K. (2002). *Acoustic and Auditory Phonetics*. 2nd Edition (1st edition, 1997). Oxford: Blackwell.

Kang, S., Johnson, K., & Finley, G. (2016). Effects of native language on compensation for coarticulation. *Speech Communication, 77*, 84-100.

Katz, W. & Bharadwaj, S. (2001). Coarticulation in fricative–vowel syllables produced by children and adults: A preliminary report. *Clinical Linguistics & Phonetics, 15*, 139–143.

Keating, P. A. (1990). The window model of coarticulation: articulatory evidence. In J. Kingston, and M. Beckman (eds.), *Papers in Laboratory Phonology I* (pp. 451-470). Cambridge: Cambridge University Press.

Kingston, J., Kawahara, S., Mash, D., & Chambless, D. (2011). Auditory contrast versus compensation for coarticulation: Data from Japanese and English listeners. *Language & Speech, 54*, 499-525.

Klatt, D. (1980). Software for a Cascade/Parallel Synthesizer. *The Journal of the Acoustical Society of America, 67*, 971-995.

Kuhl, P. K. (1991). Human adults and human infants show a “perceptual magnet effect” for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics, 50*, 93-107.

Kunisaki, O. & Fujisaki, H. (1977). On the influence of context upon perception of voiceless fricative consonants. *Annual Bulletin of the Research Institute of Logopedics & Phoniatrics, 11*, 85-91.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review, 74(6)*, 431-461.

Liberman, A. M., Delattre, P. C., Cooper, F. S., & Gerstman, L. J. (1954). The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs: General and Applied, 68(8)*, 1-13.

Liberman, A. M. & Mattingly, I. G. (1985). The Motor Theory of Speech Perception Revised. *Cognition, 21*, 1-36.

Lotto, A. J. & Kluender, K. R. (1998). General contrast effects of speech perception: Effect of preceding liquid on stop consonant identification. *Perception & Psychophysics, 60*, 602–619.

Lotto, A. J., Kluender, K. R., & Holt, L. L. (1997). Perceptual compensation for coarticulation by Japanese quail (Coturnix coturnix japonica*). The Journal of the Acoustical Society of America, 102*, 1134-1140.

Mann, V. A. (1980). Influence of preceding liquid on stop-consonant perception. *Perception & Psychophysics, 28*, 407–412.

Mann, V. A., & Repp, B. H. (1980). Influence of vocalic context on perception of the [ʃ]-[s] distinction. *Perception & Psychophysics, 28*, 213-228.

Mann, V. A. & Repp. B. H. (1981). Influence of preceding fricative on stop consonant perception. *The Journal of the Acoustical Society of America, 69*, 548-558.

Mann, V. A., & Soli, S. (1991). Perceptual order and the effect of vocalic context on fricative perception. *Perception & Psychophysics, 49*, 399–411.

Martin, J. G. & Bunnell, H. T. (1981). Perception of anticipatory coarticulation effects. *The Journal of the Acoustical Society of America, 69*, 559-567.

Martin, J. G. & Bunnell, H. T. (1982). Perception of anticipatory coarticulation effects in vowel-stop-vowel sequences. *Journal of Experimental Psychology: Human Perception and Performance, 8*, 473-488.

Marslen-Wilson, W. & Warren, P. (1994). Levels of perceptual representation and process in lexical access: Words, phonemes, and features. *Psychological Review, 101(4)*, 653-675.

Ohala, J. J. (1981). The listener as a source of sound change. In C. Masek, R. A. Hendrick, M. F. Miller (eds.), *Papers from the Parasession on Language and Behavior* (pp. 178-203). Chicago: Chicago Linguistics Society.

Ohala, J. J. (2012). The listener a sound change: an update. In M-J Solé & D. Recasens (eds.), *The initiation of sound change: Perception, production, and social factors* (pp. 21-36). Amsterdam: John Benjamins.

Pitt, M. A. & McQueen, J. M. (1998). Is compensation for coarticulation mediated by the lexicon. *Journal of Memory and Language, 39*, 347-370.

Soli, S. (1981). Second formants in fricatives: acoustic consequences of fricative-vowel coarticulation. *The Journal of the Acoustical Society of America, 70*, 976-984.

Takagi, N. & Mann, V. A. (1995). The limits of extended naturalistic exposure on the perceptual mastery of English /r/ and /l/ by adult Japanese learners of English. *Applied Psycholinguistics, 16*, 379-405.

Viswanathan, N., Magnuson, J. S., & Fowler, C. A. (2010). Compensation for coarticulation: Disentangling auditory and gestural theories of perception of coarticulatory effects in speech. *Journal of Experimental Psychology: Human Perception & Performance, 36(4)*, 1005–1015.

Warren, P. & Marslen-Wilson, W. D. (1987). Continuous uptake of acoustic cues in spoken words recognition. *Perception & Psychophysics, 41*, 262-275.

Whalen, D. H. (1981). Effects of vocalic formant transitions and vowel quality on the /s-ʃ/ boundary. *The Journal of the Acoustical Society of America, 69*, 275–282.

Whalen, D. H. (1983). Vowel information in postvocalic fricative noises. *Language and Speech, 26*, 91-100.

Whalen, D. H. (1984). Psychological mismatches slow phonetic judgements. *Perception & Psychophysics, 35*, 49-64.

Whalen, D. H. (1989). Vowel and consonant judgements are not independent when cued by the same information. *Perception & Psychophysics, 46*, 284–292.

Whalen, D. H. (1991). Subcategorical phonetic mismatches slow phonetic judgements. *Perception & Psychophysics, 35*, 49-64.

White, K. S., Yee, E., Blumstein, S. E. & Morgan, J. L. (2013). Adults show less sensitivity to phonetic detail in unfamiliar words, too. *Journal of Memory and Language, 68*, 362-378.

Yeni-Komshian, G. & Soli, S. (1981). Recognition of vowels from information in fricatives: perceptual evidence of fricative-vowel coarticulation. *The Journal of the Acoustical Society of America, 70*, 966-975.