Theses and Dissertations

Summer 2011

Perceiving speech in context: compensation for contextual variability during acoustic cue encoding and categorization

Joseph Christopher Toscano
University of Iowa

PERCEIVING SPEECH IN CONTEXT: COMPENSATION FOR CONTEXTUAL
VARIABILITY DURING ACOUSTIC CUE ENCODING AND CATEGORIZATION

by

Joseph Christopher Toscano

<u>An Abstract</u>

Of a thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy
degree in Psychology
in the Graduate College of
The University of Iowa

July 2011

Thesis Supervisor: Associate Professor Bob McMurray

ABSTRACT

Several fundamental questions about speech perception concern how listeners understand spoken language despite considerable variability in speech sounds across different contexts (the problem of lack of invariance in speech). This contextual variability is caused by several factors, including differences between individual talkers' voices, variation in speaking rate, and effects of coarticulatory context. A number of models have been proposed to describe how the speech system handles differences across contexts. Critically, these models make different predictions about (1) whether contextual variability is handled at the level of acoustic cue encoding or categorization, (2) whether it is driven by feedback from category-level processes or interactions between cues, and (3) whether listeners discard fine-grained acoustic information to compensate for contextual variability.

Separating the effects of cue- and category-level processing has been difficult because behavioral measures tap processes that occur well after initial cue encoding and are influenced by task demands and linguistic information. Recently, we have used the event-related brain potential (ERP) technique to examine cue encoding and online categorization. Specifically, we have looked at differences in the auditory N1 as a measure of acoustic cue encoding and the P3 as a measure of categorization. This allows us to examine multiple levels of processing during speech perception and can provide a useful tool for studying effects of contextual variability.

Here, I apply this approach to determine the point in processing at which context has an effect on speech perception and to examine whether acoustic cues are encoded continuously. Several types of contextual variability (talker gender, speaking rate, and coarticulation), as well as several acoustic cues (voice onset time, formant frequencies, and bandwidths), are examined in a series of experiments. The results suggest that (1) at early stages of speech processing, listeners encode continuous differences in acoustic cues, independent

of phonological categories; (2) at post-perceptual stages, fine-grained acoustic information is preserved; and (3) there is preliminary evidence that listeners encode cues relative to context via feedback from categories. These results are discussed in relation to proposed models of speech perception and sources of contextual variability.

Abstract Approved: _____
Thesis Supervisor

_____
Title and Department

_____
Date

PERCEIVING SPEECH IN CONTEXT: COMPENSATION FOR CONTEXTUAL
VARIABILITY DURING ACOUSTIC CUE ENCODING AND CATEGORIZATION

by

Joseph Christopher Toscano

A thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy
degree in Psychology
in the Graduate College of
The University of Iowa

July 2011

Thesis Supervisor: Associate Professor Bob McMurray

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

_____

PH.D. THESIS

_____

This is to certify that the Ph.D. thesis of

Joseph Christopher Toscano

has been approved by the Examining Committee for the
thesis requirement for the Doctor of Philosophy degree in
Psychology at the July 2011 graduation.

Thesis Committee: _____
Bob McMurray, Thesis Supervisor

_____
Paul Abbas

_____
Prahlad Gupta

_____
Toby Mordkoff

_____
Gregg Oden

_____
Shaun Vecera

ACKNOWLEDGEMENTS

Scientific endeavors are ultimately shaped not by individual researchers but by the interactions between them. The work presented here would not have been possible without the assistance of a number of people. First, I owe a great deal to my advisor Bob McMurray, whose own example taught me time and again how to be a successful researcher. I hope that my own work will reflect the same energy, theoretical insight, and experimental rigor that he has demonstrated in his.

I would also like to thank the members of the Mechanisms of Audio-visual Categorization Lab (MACLab) who assisted with running participants in the experiments, lent their voices for recording, and provided valuable theoretical insights into the issues presented here. In particular, I would like to thank Dan McEchron for scheduling subjects and keeping the lab running smoothly. Also, I am incredibly thankful for Lisa Vangsness' assistance with running ERP subjects and creating stimuli used in the experiments. Her time, effort, and skill were invaluable. I am also grateful to Effie Kapnoula for getting up to speed on the ERP technique so quickly and running subjects at the eleventh hour. Other members of the lab, as well as my fellow grad students in the Cognition and Perpection area and related programs, particularly Keith Apfelbaum, Ashley Farris-Trimble, Marcus Galle, Katy Mueller, and Gwyn Rost have provided me with a great deal of intellectual energy as well as opportunities to take a break every once in a while.

I would also like to acknowledge the other members of my dissertation committee, Paul Abbas, Prahlad Gupta, Toby Mordkoff, Gregg Oden, and Shaun Vecera, for their helpful comments at various stages of the dissertation process. I am especially grateful to Toby for his generosity in the use of his ERP lab, the time and effort he put into developing the software and setting up the equipment used to collect the data, invaluable comments about the design and interpretation of ERP experiments, and for reminding me of the importance

TABLE OF CONTENTS

iv

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER 1
INTRODUCTION

1.1   Lack of invariance in speech

Perceptual systems must build a model of the world from extremely noisy sensory input. This problem is particularly apparent in speech perception. Language users rely heavily on spoken language in their daily lives, and despite enormous variability in the speech signal, they are remarkably adept at correctly recognizing speech sounds. Often, listeners are not even aware of this variability, and the speech system handles it effortlessly, decoding the meaning of a spoken utterance even when speech is spoken extremely quickly, by someone they have never heard before, or when adjacent sounds blend together (as they normally do in conversational speech).

A basic question then is how listeners transform a highly variable, transient speech signal into discrete linguistic units, like phonemes or syllables, during perception. This process involves mapping continuous acoustic cues (i.e., information that distinguishes phonetic contrasts) onto categories (phonological features, phonemes, or words). As a minimal description, this can be characterized as a two-stage process: (1) encoding acoustic information and (2) categorizing that information.

The simplest way to do this would be to directly map specific cue-values onto specific phonological categories. For example, to distinguish stop consonant voicing categories (/b,d,g/ vs. /p,t,k/), listeners can use voice-onset time (VOT; the timing between the release of consonantal closure and the onset of glottal voicing), mapping short VOTs (below $\approx$20 ms in English) to voiced phonemes and long VOTs (>20 ms) to voiceless phonemes (Lisker & Abramson, 1964).

However, there simply is no one-to-one mapping between acoustic cues and linguistic categories (Repp, 1981). One of the main reasons for this is that speech sounds are heavily influenced by context. VOT, for example, varies as a function of speaking rate with talkers

producing shorter VOT-values when speaking quickly and longer VOT values when speaking slowly. Thus, a given VOT-value is context dependent, and as a result, the boundary separating the categories is also context dependent. Indeed, listeners will recognize a word with an intermediate VOT-value as /p/ 90% of the time when it is spoken in a fast sentence and as /b/ 70 % of the time when spoken slowly (Toscano & McMurray, 2011a).

This contextual variability has been described as the *problem of lack of invariance* in speech: any given acoustic cue contains information about multiple phonemes and is affected by context (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Fowler, 1984). Similarly, any given phoneme is determined by multiple acoustic cues (Repp, 1982; Lisker, 1986; Jongman, Wayland, & Wong, 2000). Moreover, cues can be temporally asynchronous within a word and can span adjacent words with intervening phonemes (Fowler, 1984). Thus, speech perception is organized around context-dependent cue-to-category mappings that must be integrated over time. Understanding how listeners do this is critical for understanding language processing and for creating systems designed to accurately recognize speech.

1.2 Approaches to handling contextual variability

Research in speech perception over the last 60 years has focused a great deal of effort on understanding three general sources of contextual variability: speaking rate, individual talker differences, and coarticulatory context. A number of approaches for handling each type of variability have been proposed. These proposals range from the idea that listeners encode acoustic cues relative to context (Kluender, 2003; McMurray & Jongman, 2011; Ladefoged & Broadbent, 1957; Lobanov, 1971), to those that propose that they take context into account at later stages of processing (Oden & Massaro, 1978; Nearey, 1990; Smits, 2001b), to those that suggest they rely on a small number of context-invariant cues (Stevens & Blumstein, 1978; Syrdal & Gopal, 1986; Summerfield, 1981), to those that suggest they rely on a massive number of context-sensitive cues (Goldinger, 1998; Johnson, 1997).

All of these models generally agree that listeners must map continuous cue-values onto categories and that context must be somehow taken into account. However, there are several critical distinctions between different approaches that rest at the level of cues themselves. Are cues encoded independently of context, with listeners taking context information into account at a later stage of processing (e.g., during phoneme categorization)? Or, is cue encoding context-relative, such that listeners encode cues differently in different contexts? If so, is this process driven directly by the acoustic properties of the context, or by feedback from more abstract, category-level information?

These basic questions apply to all types of contextual variability, whether listeners are compensating for speaking rate, talker-specific factors, or coarticulation. However, some approaches make different predictions for different types of context effects. Several proposals have been framed in terms that are specific to a particular type of effect (e.g., timing processes designed to deal with variation in rate; Summerfield, 1981), while other accounts are more general purpose (Nearey, 1990; Oden & Massaro, 1978). An open question is whether there is a single process for dealing with different sources of contextual variability or whether there are multiple mechanisms used to compensate for talker, coarticulatory, and speaking rate differences.

1.3 Measuring cue-level processing

Differentiating between these accounts requires us to assess cue-level encoding, that is, the initial process of recognizing the relevant acoustic cues before mapping them onto categories. However, this is difficult to do for speech sounds, since listeners' responses to variation in sounds along continuous phonetic dimensions are influenced by categorization (Massaro & Cohen, 1983; Gerrits & Schouten, 2004; Pisoni, 1973; Pisoni & Lazarus, 1974). Partially as a result of this, a great deal of early work suggested that speech perception is fundamentally categorical and that listeners do not encode continuous cue-values. Even discrimination tasks, which should provide the best means of assessing cue encoding

behaviorally (Pisoni & Lazarus, 1974), show effects of listeners phonological categories: listeners are better at discriminating sounds that span a category boundary than those within a boundary, even if the acoustic distance between the sounds is the same (Liberman et al., 1967). Moreover, behavioral responses occur well after initial perceptual processing, and although online measures of speech processing, such as eye movements in the visual-world task, demonstrate effects of within-category phonetic differences (McMurray, Tanenhaus, & Aslin, 2002), they reflect listeners' categories.

Given this, it is unclear whether cue encoding is fundamentally categorical or whether these measures simply do not reflect cue encoding. This is critically important since some models of context compensation propose that listeners are highly sensitive to fine-grained acoustic detail (McMurray & Jongman, 2011; Johnson, 1997; Goldinger, 1998) while others do not make such claims.

In order to obtain a better measure of cue-level processes and to examine whether cue encoding is categorical or continuous, Toscano, McMurray, Dennhardt, and Luck (2010) recently used the event-related brain potential (ERP) technique to look at electrophysiological responses to speech sounds. We examined changes in the auditory N1 and P3 components as a function of VOT and voicing category. An N1 is produced in response to a variety of auditory stimuli and occurs late in perceptual processing but early in language processing. This suggested that it may serve as a good measure of continuous cue encoding. The P3 component is related to later decision processes, and therefore, may be a good indicator of categorization. Using auditory stimuli varying in VOT, we found that the amplitude of the N1 varies linearly with changes in VOT and is not influenced by phonological categories or listeners' responses. The later-occurring P3 component also varies with VOT but depends on the phonological category the listener is assigning the stimuli to (either voiced [e.g., *beach*] or voiceless [*peach*]).

Two important conclusions can be drawn from these results. First, during cue en-

coding, listeners are sensitive to acoustic differences in a way that is not related to phono-logical categories (as seen in the N1 response). In addition, within-category sensitivity is maintained even at post-perceptual stages (as seen in the P3). These results argue strongly against categorical perception and indicate that listeners are highly sensitive to fine-grained acoustic information within a single phonetic category. This, in itself, has implications for how listeners handle contextual variability, since a number of approaches have argued that listeners discard acoustic information when compensating for contextual differences (see Pisoni, 1993, for a discussion of this).

Second, these results suggest that we may be able to use the N1 as an index of cue-level perceptual encoding, independent of the phonological categories of the language. Thus, this technique may be useful for more closely assessing questions about contextual variability in speech perception. By examining cue encoding, we can differentiate between different models of context compensation in a way that does not seem to be possible with behavioral measures alone.

The experiments presented here follow the approach of Toscano et al. (2010), using the auditory N1 to measure cue encoding and the P3 component to measure categorization, in order to examine how listeners handle contextual variability. This is done using a variety of classic context effects in speech perception, looking at contextual information from speaking rate, talker gender, and neighboring phonemes (coarticulatory information).

1.4 Overview of dissertation

The present study evaluates a number of models of context compensation in speech perception using the ERP paradigm developed by Toscano et al. (2010). In doing so, these experiments provide additional details about what information is indexed by auditory N1 responses to speech sounds by looking at new phonological contrasts and acoustic cues.

Specifically, I examined the effect of preceding context on perceptual encoding (as measured by the auditory N1) and categorization (measured by the P3) for three different

types of contextual variation: (1) talker gender, (2) speaking rate, and (3) coarticulatory context. Each of these is a major source of the lack of invariance in speech, and, simultaneously, each provides a predictable (though different) source of information that listeners can use to compensate for differences across contexts. In addition, several acoustic cues were used to look at these context effects, specifically, VOT and formant (F1, F2, and F3) frequencies and bandwidths, allowing us to assess whether they are encoded continuously.

The goals of this study are to determine (1) whether cues are encoded relative to preceding context information in a way that preserves fine-grained acoustic detail, (2) whether feedback from higher-level stages affects cue encoding, and (3) whether these effects apply similarly to different types of context information. Together, this will allow us to uncover the principles listeners use to overcome contextual variability in speech.

The organization of the rest of this dissertation is as follows. The next chapter presents a core set of empirical findings describing context effects in speech perception and discusses the proposed accounts of how listeners deal with contextual variability. Chapter 3 addresses some general methodological issues that apply to the ERP technique used here and two experiments examining N1 responses to the speech sounds used in subsequent experiments.

Chapters 4-6 introduce each of the three sources of contextual variability and the experiments designed to measure their effects. Each starts with a brief overview of the more-specific literature on the source of context information in question and then describes a series of experiments examining its effects. Chapter 4 focuses on talker variability, Chapter 5 examines speaking rate variability, and Chapter 6 addresses coarticulatory effects. Finally, Chapter 7 presents a summary and discussion of the results of the experiments.

CHAPTER 2
CONTEXTUAL VARIABILITY IN SPEECH

Speech sounds are affected by many sources of contextual variability. Differences in speaking rate, talkers' voices, and neighboring phonemes all produces changes in phonetic cues in a given segment such that listeners must compensate for them during speech recognition in order to correctly identify that segment. A number of models have been proposed for handling particular context effects or as more general solutions to the problem of lack of invariance. These models differ in two important respects that will be discussed in detail below: (1) whether raw cue-values or context-relative cues are encoded during perception, and (2) the direction of information flow (e.g., feedforward vs. feedback) that drives context compensation.

Here, I will review each of the accounts that have been proposed. First, I will present a brief overview of the basic phenomena related to the three types of contextual variability that will be examined in the experiments — speaking rate, talker, and coarticulatory information. Second, I will describe the two features that differentiate models and the predictions that they make about how listeners handle variability across contexts.

2.1 Sources of contextual variability

A brief overview of each source of contextual variability is given here in order to introduce them. A more detailed review, particularly focusing on the effects examined in a specific set of experiments, are given in Chapters 4 (talker variability), 5 (speaking rate), and 6 (coarticulation).

2.1.1 Talker identity

A number of acoustic cues in speech are influenced by differences between talkers' voices, and several studies have specifically explored the effects of talker gender, which produces one of the most salient differences. Variation in acoustic cues between men and

women occur both because of differences in the size and shape of the articulators (Peterson & Barney, 1952; Hillenbrand, Getty, Clark, & Wheeler, 1995) and for sociophonetic reasons (Strand & Johnson, 1996).

One well-studied example of this occurs in vowel production. Peterson and Barney (1952) collected a large set of phonetic data on formant frequencies for vowels spoken by a larger number of men, women, and children. They found considerable overlap between vowel categories signaled by the frequencies of the first two formants (F1 and F2) and large differences in the size and shape of the vowel spaces for the three groups, such that formant frequencies for women were higher than those for men, consistent with the differences in the average size of the vocal tract for each group. These differences between talkers contribute to a great deal of the variability seen in F1 and F2 values. Illustrating the importance of this Cole, Linebaugh, Munson, and McMurray (2010) found that approximately 80% of the variance in F1 and 40% of the variance in F2 in a corpus of /ʌ/ and /ɛ/ measurements was attributable to talker-specific factors.

There are also phonetic differences between talkers that are due to learned sociophonetic variability. Women have higher spectral means than men for frication sounds, which distinguishes fricative place of articulation in sibilant fricatives (Strand & Johnson, 1996), though it does not seem that this effect is derived from physiological differences in the vocal apparatus, suggesting that these effects may be socio-phonetically driven. As with differences that arise because of articulatory factors, listeners would need to compensate for this variability in order to use spectral mean as a cue to fricative place, though models of talker compensation may handle articulatory and sociophonetic effects differently.

Listeners are also sensitive to these phonetic differences. Ladefoged and Broadbent (1957) presented early evidence for this, demonstrating that listeners label vowels differently as a function of talker-specific information in a preceding carrier phrase. They presented listeners with the sentence "Please say what this word is" followed by a word of the form

/bVt/ (e.g., *bit*, *bat*, etc.). Six versions of the sentence were synthesized by adjusting its formant frequencies, creating the perception of six different talkers. They found that listeners' categorized a given vowel sound differently depending on the formant-values in the preceding sentence. For example, if a sentence had lower overall F1 values, listeners treated F1 in the target word as higher than it actually was.

Another approach for examining the perceptual consequences of talker variability (as well as other context effects) is to vary a phonetic cue along a continuous dimension (e.g., F1 as a cue to vowel quality) and examine shifts in listeners' category boundaries along that dimension as a function of context. Johnson, Strand, and D'Imperio (1999) used this technique to examine listeners' responses to stimuli that varied along an F1 continuum from *hood* to *hud* and were spoken by either a male or female talker. They found that listeners' category boundaries shifted such that the boundary was lower for men (i.e., listeners made more "hud" responses). Thus, listeners compensate for differences between talker's voices in a way that is consistent with the phonetic data.

Together, the phonetic and perceptual data on talker gender differences suggest that listeners handle variability between men and women's voices that affects a variety of acoustic cues (formant frequencies, VOT, and spectral mean) and phonological features (vowel quality, voicing, and place of articulation).

2.1.2   Speaking rate

Speaking rate, which can be signaled either by sentential rate (SR) in running speech, or by vowel length (VL) in isolated words (though VL may actually be a weak phonetic cue; see Toscano & McMurray, 2010, 2011a) is another source of contextual variability. This also has an effect on several phonological distinctions, particularly contrasts signaled at least partially by temporal cues. These include vowel quality, voicing, and manner of articulation (Miller, 1981). Because some acoustic cues to these distinctions are defined as time differences (e.g., VOT, formant transition durations), cue-values vary as a function of

speaking rate. Thus, as with talker variability, listeners must compensate for rate differences to accurately use these cues.

Phonetic data on VOT values spoken at different rates illustrates this effect. Recall that VOTs near 0 ms typically indicate voiced sounds in English, while longer VOTs (50 ms) indicate voiceless sounds. Allen and Miller (1999) found that, ignoring voicing category, talkers produced longer VOT values at slower speaking rates, and shorter VOTs at faster rates. Thus, VOT can be ambiguous when used as a cue by itself — a VOT of 25 ms may indicate a voiceless sound in fast speech, and a voiced sound in slow speech.[1]

Perceptual data show that listeners are sensitive to these differences. Summerfield (1981) presented listeners with synthetic speech that started with either a voiced (/b/) or voiceless (/p/) stop. He varied both the VOT of the stimuli and the length of the subsequent vowel and found that vowel length shifted the category boundary between voiced and voiceless sounds, such that for longer vowels (i.e., slower speech), more sounds were perceived as voiced, and for shorter vowels, more sounds were perceived as voiceless. Other studies have found effects of the preceding SR (Miller & Grosjean, 1981; Summerfield, 1981), and similar effects of VL for listeners' manner of articulation judgments based on formant transition duration (Miller & Liberman, 1979).

Together, these results suggest that listeners compensate for rate differences indicated by both SR and VL, similar to the compensation seen for talker gender differences.

2.1.3 Coarticulation

A third source of contextual variability in speech perception arises from coarticulation. In running speech, the acoustic properties of most phonemes are heavily influenced by coarticulation from surrounding phonemes (Liberman, Delattre, & Cooper, 1952), since, at any point in time, the position of the articulators is influenced by both past and future

---

[1]This effect is driven by differences in VOT values of voiceless stops, which have longer VOTs than voiced stops and more within-category variability.

segments (Fowler, 1984). As with talker and rate variability, listeners must compensate for these differences to correctly identify speech sounds.

Coarticulatory effects occur for a large number of cues and phonological distinctions in speech. One well-studied example is the effect of vowel context on stop consonant place of articulation (/b,p/ vs. /d,t/ vs. /g,k/). Kewley-Port (1982) presented an extensive acoustic analysis of the properties of voiced stop consonants (/b,d,g/) produced in the context of different vowels. She found that formant frequency transitions, despite being good cues to place of articulation, were not independent of vowel context. For example, /d/ generally has a lower F1 onset than /b/. However, F1 onset frequency for the syllable /bi/ is actually much *lower* (310 Hz) than for the syllable /da/ (392 Hz). Thus, raw F1 onset frequency cannot be used as a cue to place of articulation across different vowel contexts. In order to use formant transitions to distinguish stops, listeners must either compensate for these differences, or compute formant onset frequencies relative to the formant frequency of the neighboring vowel (i.e., use *locus equations*), which would presumably be invariant across contexts (Sussman & Shore, 1996).

Listeners also compensate for these coarticulatory differences during perception (Repp & Mann, 1981; Mann & Repp, 1981; Liberman et al., 1952). Liberman et al. (1952) provides one of the earliest demonstrations of this. They presented listeners with synthetic consonant-vowel syllables generated by noise bursts followed by periodic energy, producing an initial sound was heard as a voiceless stop consonant (/p,t,k/). They found that listeners' perception of the place of articulation of the initial consonant depended on not only the frequency of the noise burst, but also the frequency of the periodic energy. That is, listeners identified stimuli with the same noise burst differently depending on the subsequent vocalic information (see also Repp & Mann, 1981; Mann & Repp, 1981).

These results demonstrate that, as with talker and rate variability, listeners adjust for coarticulation and alter phonetic judgments accordingly during speech perception.

2.1.4 Differences between sources of contextual variability

While each of these contextual factors produces predictable variation in phonetic cues, they differ in the nature of the information about the context that is available to listeners. For example, talker identity is more-or-less categorical, while variation in speaking rate is more continuous and does not support a fixed number of discrete categories. In addition, coarticulation differs from the other two in that it reflects categorical information (the identity of the neighboring phoneme) but is bidirectional (each phoneme serves as context for the other).

In addition, these three types of context differ in when the relevant context information is available. Rate and talker information can be signaled by acoustic information that occurs well-before the segment that must undergo compensation, as well as well-after. These have been described as remote context effects (Repp, 1982). In contrast, the effects of coarticulation are most evident for adjacent segments and have been described as local context effects. However, the distinction between remote and local context effects is not always clear, since effects of coarticulation can span multiple phonemes (Alfonso & Baer, 1982), as well as word boundaries (Fowler, 1984; Cole et al., 2010), and rate effects can be seen in isolated words (Miller & Liberman, 1979). A number of models that have proposed cue-level context compensation differ depending on whether listeners can take advantage of long-distance context information or only short-range information (Nearey, 1989).

Context information can also potentially affect cue encoding in other ways. For example, Bradlow, Toretta, and Pisoni (1995) examined effects of speech intelligibility as a function of several characteristics, including talker gender and speaking rate. They found an overall effect of gender, such that women were more intelligible than men. In contrast, they did not find an effect of speaking rate; fast speech had a similar level of intelligibility to slow speech. Thus, listeners may process speech in talker contexts differently because of differences in intelligibility, rather than as a form of compensation for talker-specific

differences. This, in turn, may have an effect on cue encoding that is unrelated to context compensation (e.g., listeners may do more work to process speech from one talker than another in order to handle differences in speech intelligibility).

As discussed below, these differences are particularly relevant, since some of the proposed explanations for how listeners handle contextual variation differ in both what information listeners need in order to compensate for context differences at different levels of processing and whether the mechanisms used to handle this variability would apply similarly to all sources of context information.

2.2 Solutions to the problem of contextual variability

Several solutions for handling each of these effects, as well as several models aimed at explaining contextual variability in general, have been proposed. Figure 2.1 shows schematics of several types of models. There are two main characteristics that distinguish the different approaches: (1) whether cues are encoded in relation to context information or whether listeners use raw cue-values, and (2) whether compensation (either at the level of cue encoding or categorization) is performed via completely feedforward processes, via feedback, or through lateral interactions. An overview of both of these characteristics and examples of models illustrating them are given below. (A detailed discussion of specific models applied to each of the three types of contextual variability is given in Chapters 4-6.)

Note that not all approaches to context compensation fit neatly into a specific category — some models have characteristics of multiple categories along the two dimensions. For example, the computing cues relative to expectations (C-CuRE) approach (McMurray & Jongman, 2011) uses both relative and raw-cue encoding. Similarly, some models do not emphasize differences along these two dimensions. Gestural parsing approaches (Fowler, 1984), for example, focus on the idea of overlapping articulatory gestures and don't fit clearly into this classification system. Thus, these two dimensions (type of encoding and direction of information flow) are simply meant to capture the key properties of the majority of speech

perception models that have been used to explain compensation for contextual variability.

2.2.1 Type of cue encoding

The first of the two characteristics concerns whether listeners encode acoustic cues relative to context or whether they encode raw cue-values, representing the incoming speech signal veridically. Here, a "cue" refers to aspects of the speech signal that can be directly measured (e.g., the steady-state frequency of the first formant).

Most models can generally be divided along this dimension based on whether they argue for *intrinsic*, *extrinsic*, or *raw* cue encoding (corresponding to different columns in Figure 2.1). Intrinsic approaches suggest that listeners encode cues in relation to each other within a given segment (Syrdal & Gopal, 1986; Nearey, 1989; Christovich & Lublinskaya, 1979; Miller & Liberman, 1979; Summerfield, 1981; Fujisaki & Kawashima, 1968). In contrast, extrinsic approaches argue that listeners encode cues in relation to the broader context, that is, that they use more abstract information or information from surrounding segments to encode cues (Nearey, 1989; Ladefoged & Broadbent, 1957; Lotto & Kluender, 1998; Lobanov, 1971). Finally, a third class of models suggests that listeners do not encode cues in relation to context at all, but rather, that they encode raw cues and compensate for context effects at later stages of processing (Smits, 2001b; Toscano & McMurray, 2010; Nearey, 1997). Note that there are some models that combine aspects of both relative and raw-cue encoding approaches (Oden & Massaro, 1978; Cole et al., 2010; McMurray & Jongman, 2011). The way that both raw and relative cues are used in these models is discussed below.

2.2.1.1 Intrinsic encoding

One of the earliest solutions to the problem of contextual variability was centered on the idea that listeners simply use information that is invariant across contexts. Many models using this approach have been referred to as intrinsic (or compound-cue) models, since they argue that listeners handle contextual variability using a small number of cues

	Type of encoding		
	Intrinsic (relative)	Extrinsic (relative)	Raw cues
Lateral	Invariance and cue-ratios (Stevens, Blumstein, Sussman, Syrdal)	Auditory contrast (Diehl, Holt, Kluender, Lotto)	Diphone biases (Nearey)
Feedforward	n/a	n/a	Cue integration (Smits; Toscano & McMurray)
Feedback	n/a	Computing cues relative to expectations (McMurray et al.)	Interactive activation (McClelland & Elman)

Figure 2.1: Schematic illustrations of models for context compensation based on their information processing pathways. Columns correspond to different types of cue encoding (extrinsic, intrinsic, or raw-cues), and rows correspond to different approaches for how context information and cues interact (lateral, feedforward, or feedback processes). Examples of specific models that fall under each category are listed below the model schematics.

that are defined as relationships between auditory information within a phonetic segment (left column in Figure 2.1). Thus, intrinsic models suggest that cues are encoded relative to context at the earliest stages of perception. Note that, in some models (e.g., the fuzzy logical model of perception [FLMP]; Oden & Massaro, 1978), the input may be defined either in terms of raw acoustic cues (see below) or in terms of intrinsic relationships between cues.

The need for compound cues was initially motivated by the lack of context-invariance for many individual acoustic cues. There is empirical support for context-invariant cues for at least some phonetic contrasts (Stevens & Blumstein, 1978; Stevens, 2002; Stevens & Keyser, 2010; Port & Dalby, 1982). For example, Stevens and Blumstein (1978) presented early evidence suggesting that spectral shape serves as an invariant cue that allows listeners to identify place of articulation (e.g., /b/ vs. /d/ vs. /g/) across different vowel coarticulatory contexts. However, many researchers have concluded that a sufficient number of invariant cues does not exist for all phonological distinctions (e.g., Ohala, 1996; Lindblom, 1996; McMurray & Jongman, 2011).

Other researchers have defined invariant cues in terms of relationships between multiple, context-sensitive cues (Sussman, Fruchter, Hilbert, & Sirosh, 1998; Sussman & Shore, 1996; Boucher, 2002; Pind, 1995). For example, Sussman and Shore (1996) suggested that locus equations (defined by formant transition slope and intercept) are sufficient for identifying the place of articulation of a sound across different manners of articulation (stops vs. nasals), voicing, and vowel contexts, effectively compensating for coarticulatory effects. These compound cues represent a form of intrinsic encoding in which contextual variability is handled via relationships between acoustic information within the segment or syllable.

To use this approach to compensate for speaking rate differences, for example, listeners could use a compound cue that is more invariant, such as the ratio between VOT (or closure duration for word-final consonants) and vowel length, rather than using VOT itself. A lower VOT:VL ratio would be more indicative of a voiced sound, and a higher ratio

would be indicative of a voiceless sound (Boucher, 2002). Talker variability provides another well-known example of the use of intrinsic models. Several researchers have proposed solutions that handle talker variability by relying on differences between formants (Syrdal & Gopal, 1986) or formant ratios (Fujisaki & Kawashima, 1968). As with the use of VOT:VL ratios to create rate-invariant cues, formant relationships provide a source of information for intrinsic compensation of talker variability.

To summarize, intrinsic encoding models all predict that listeners encode cues relative to the immediate context, using information within a given phonetic segment to compensate for contextual variability.

2.2.1.2 *Extrinsic encoding*

Extrinsic approaches (center column in Figure 2.1) offer another class of solutions for taking context into account during cue encoding. In contrast to intrinsic models, extrinsic compensation models suggest that listeners compute cues relative to information outside the immediate syllable, rather than via relationships between phonetic cues within a syllable (though some researchers have argued for combining intrinsic and extrinsic approaches, e.g., Nearey, 1989).[2] Long-distance differences across context produce predictable changes in upcoming acoustic cue-values (J. G. Martin & Bunnell, 1981). Thus, the effect of preceding context can be factored out of the cue, leaving a less context-sensitive source of information available for speech recognition.

Using the example of speaking rate again, extrinsic compensation processes could be used to compute VOT relative to the preceding sentence rate (Dupoux & Green, 1997). At faster rates, VOT values tend to be shorter. Thus, a listener could recode a given VOT

---

[2]Here, I am referring to intrinsic and extrinsic compensation *mechanisms*, that is, ways contextual variability can be handled at the level of cue encoding. The distinction between intrinsic and extrinsic *sources of information* is also relevant, and these concepts could be applied to raw-cue encoding approaches as well. However, since most raw-cue models do not distinguish between the two sources of information, the focus here is on intrinsic and extrinsic effects as mechanisms for relative cue encoding.

as longer than it actually is in order to compensate. Similarly, at slow rates, that same VOT would be recoded as shorter. This has a similar effect to that of VOT:VL ratios used in intrinsic approaches: listeners use relative cue-values, rather than raw ones, to make voicing judgments independent of context.

Extrinsic approaches have also been widely applied to coarticulatory context effects. Fowler's (1984) gestural parsing account seems to fit within this category (since it suggests listeners use information from adjacent segments), though it does not emphasize the distinction between intrinsic and extrinsic approaches. Gestural parsing models suggest that speech is encoded in terms of a series of articulatory gestures for which differences due to coarticulation are factored out. Although this differs from the other models discussed so far in that it suggests speech is processed in terms of articulatory gestures rather than acoustic cues, the same principle of extrinsic compensation applies here as well. That is, in a gestural parsing account, listeners use information from adjacent segments to recode the incoming acoustic information as context-invariant gestures.

Other researchers have suggested similar processes that operate over acoustic cues rather than gestures. These models emphasize general auditory principles, rather than speech-specific ones, that allow listeners to compensate for contextual variability (Pisoni, Carrell, & Gans, 1983; Diehl & Kluender, 1989; Lotto & Kluender, 1998; Kluender, 2003; Holt, 2005). These approaches suggest that listeners encode sounds relative to preceding context in a contrastive way (i.e., sounds are encoded as a lower frequency in the context of a preceding high-frequency sound; they are encoded as shorter in the context of a preceding sound that is longer). For example, third formant frequency (which provides a cue to the /d/-/g/ distinction) would be coded as higher if it follows a low frequency sound and lower if it follows a high frequency sound (Lotto & Kluender, 1998). This approach is based on basic auditory principles, and is not speech-specific — a preceding pure-tone, for example, is sufficient to alter speech perception (Lotto & Kluender, 1998; Holt, 2005).

Along with intrinsic approaches, extrinsic compensation approaches have typically been used as *normalization* procedures that discard fine-grained acoustic information to deal with contextual variability. However, listeners do not seem to do this when compensating for context differences (Pisoni, 1993). Some models allow for the possibility of extrinsic encoding while preserving fine-grained acoustic detail (e.g., FLMP; Massaro & Oden, 1980). Recently, McMurray and colleagues (Cole et al., 2010; McMurray, Cole, & Munson, in press; McMurray & Jongman, 2011) have proposed a specific computational framework for cue compensation — computing cues relative to expectations (C-CuRE) — that fits within the class of models that use extrinsic encoding, as well as raw cues (see below), and argues that listeners do not discard information when compensating for contextual variability.

In C-CuRE, cues are initially encoded as raw cue-values, which are heavily influenced by context. Then, as sources of context information are received, listeners set up expectations for particular cue-values. This allows them to factor out variation due to context from the raw cues via processes similar to linear regression. The residual (the difference between the actual cue-value and its expected value after context has been taken into account) is then used as a new, less context-variable cue. This is similar to the auditory contrast approach in that cues are computed relative to preceding context, and it is similar to gestural parsing in that contextual variation is factored out of cue-values on the basis of expectations about particular segments. Moreover, it suggests that listeners handle context differences on the basis of the relationship between cues and context information, which may not necessarily be contrastive (as suggested by auditory contrast models). This approach is computationally very powerful and can greatly reduce variance due to both coarticulation (Cole et al., 2010; McMurray et al., in press) and talker variability (McMurray & Jongman, 2011).

In general, extrinsic encoding models all predict that listeners encode cues relative to context. This is similar to intrinsic encoding models, but unlike those models, extrinsic

approaches suggest that listeners use information from preceding and adjacent segments rather than context information contained within a single segment.

2.2.1.3   *Raw-cue encoding*

The previous two approaches both suggest that cue-values are coded as relative values. In contrast to this, *raw-cue* models (right column in Figure 2.1) suggest that listeners do not compute relative cue-values at all. Rather, they argue that context compensation occurs at later stages of processing. Thus, the distinction between extrinsic/intrinsic models and raw-cue models is made apparent at the level of cue encoding.

Models that suggest listeners encode raw cue-values address the problem of contextual variability by using context information as input to higher-level representations, such as phoneme or lexical categories. This approach effectively treats contextual variability similarly to variability within typically-defined phonetic cues like VOT: by mapping context information onto a specific category. For example, long VLs could be mapped onto the /b/ category, and short VLs onto the /p/ category, just as short VOTs are mapped to /b/ and long VOTs are mapped to /p/ (Toscano & McMurray, 2010). Thus, although VL is typically described as an indicator of speaking rate, it can also serve as a weak phonetic cue for word-initial voicing distinctions. Indeed, Allen and Miller's (1999) phonetic data showing different VL distributions for the two voicing categories shows that VL carries information about voicing categories.

An advantage of these approaches is that they can be applied using the same learning principles and categorization processes used to handle multiple phonetic cues (Toscano & McMurray, 2010). Moreover, they fit well with more general cue-integration models that describe how listeners use multiple acoustic cues for a given phonological distinction (Nearey, 1997, 1990, 1986; Oden & Massaro, 1978; Toscano & McMurray, 2010; Smits, 2001b). Cue-integration models, such as hierarchical categorization (HICAT; Smits, 2001b) and weighted Gaussian mixture models (WGMM; Toscano & McMurray, 2010), handle contextual vari-

ability by encoding raw cues and compensating for context at later stages of processing. Nearey and colleagues' normalized *a posteriori* probability (NAPP) models (Nearey, 1997, 1990, 1986) use raw cue-values to encode the incoming stimulus and take context into account at the level of diphones. FLMP shares properties with both raw and relative encoding approaches: it can use information from raw acoustic cues (Oden & Massaro, 1978), but it also uses inputs that are normalized based on category prototypes (Massaro & Cohen, 1976) and can use relative cue encoding (Massaro & Oden, 1980).

These models differ in the level at which context information is combined with phonetic categories, with some suggesting this occurs at the level of phonological features (WGMM; Toscano & McMurray, 2010), others suggesting it happens at the level of diphones, where weak biases between pairs of phonemes allow listeners to show apparent context effects when the information for one phoneme is ambiguous (NAPP; Nearey, 1997, 1990), and other suggesting that boundaries along the cue dimension shift as a function of context (HICAT; Smits, 2001b).

The number of cues used can also vary, with exemplar models (Johnson, 1997; Goldinger, 1998) representing an extreme form of this type of model. In exemplar models, listeners store phonological categories (or words) as a large number of context-sensitive cues. Crucially, cues for contextual factors, such as talker and rate, are stored along with phonetically-contrastive cues. This means that they can also be used to activate exemplars (of phonemes or words) as a way of handling contextual variability.

In general, raw-cue encoding fits with the notion that both multiple cues and context effects produce shifts in listeners' categorization functions along a cue dimension (illustrated in the examples in Section 2.1). However, a number of researchers have distinguished between trading relations (i.e., shifts in listeners categorization responses producing equivalent perceptual judgments for combinations of multiple cues) and context effects. In his review of trading relations and context effects in speech perception, Repp (1982) makes the

distinction between acoustic events caused by the way a particular speech sound is produced (cues) from events that are not (context effects). He uses an example from Mann and Repp (1980) in which listeners' judgment of fricative place (/ʃ/ vs. /s/) is affected by both formant transitions and the identity of the vowel in the vocalic portion. Here, the effect of formant transitions would be classified as a trading relation (since they are produced as a consequence of which fricative the talker intended) and would not affect listeners' representation of other cues, consistent with a raw-cue encoding approach. In contrast, the effect of vowel identity would be a context effect (since it is not a consequence of which fricative the talker intended) and would influence encoding of other cues, consistent with extrinsic encoding. Raw-cue approaches to handling contextual variability challenge this distinction, suggesting that context information does not affect listeners encoding of acoustic cues.

Overall, raw-cue models predict that listeners do not compensate for contextual variability at the level of cue encoding. Thus, this distinguishes them from the previous two approaches in that they suggest context information is taken into account at later stages of processing and that cue encoding should be the same in different contexts.

2.2.2 Direction of information flow

In addition to the distinction between relative encoding (intrinsic and extrinsic) and raw-cue encoding, approaches to handling contextual variability can also be distinguished in the basis of how context information is combined with phonetic cues (corresponding to the different rows in Figure 2.1). This can generally be implemented in three ways: (1) lateral relationships at a single processing level between phonetic cues and cues related to context, (2) feedback from higher-level category representations onto cues, and (3) completely feedforward systems where lower-level context information affects phoneme (or similar) category-level representations. The characteristics of models that use feedforward, lateral, and feedback processes are discussed below.

Note that the type of encoding and direction of information flow are not perfectly

separable dimensions. Intrinsic encoding implies lateral relationships, since context compensation happens at the level of cue encoding via cue-cue interactions. Similarly, feedforward models necessarily all use raw-cue encoding, since no lateral or feedback processes can be used to compute relative cues in those models. In addition, as with the type of encoding, some models do not fit neatly into a specific category along this dimension.

2.2.2.1 Lateral models

Lateral models (top row in Figure 2.1) generally suggest that acoustic information from context is combined with phonetic cues at the level of cue encoding. All models of intrinsic encoding are inherently lateral models (Syrdal & Gopal, 1986; Summerfield, 1981; Fujisaki & Kawashima, 1968; Sussman & Shore, 1996). In these models, multiple sources of information are combined at the level of phonetic cues; no interaction with other levels of processing, such as phoneme or indexical categories, is needed. In Syrdal and Gopal's (1986) model of talker compensation, context-independent vowel cues are computed as differences between formant and pitch values. This process occurs entirely at the level of cue encoding. Similarly, Summerfield's (Summerfield, 1981) rate compensation model suggests that rate-independent VOT estimates are computed from continuous estimates of VL. Rate compensation approaches that subtract a continuous estimate of preceding sentence rate from VOT values would also fit into this category.

Several extrinsic encoding approaches also use lateral interactions to compensate for contextual variability. In this type of model, cues are recoded via extrinsic information, since listeners are compensating for context effects at the level of cue encoding using information from adjacent segments. The auditory contrast account of coarticulatory compensation (Lotto & Kluender, 1998) is a clear example of this, since it factors out information from preceding segments via interactions between different cues.[3]

---

[3]Gestural parsing accounts (Fowler, 1984) may also fall under this category, but it is not clear whether coarticulatory effects are handled via direct interactions between continuous information

While raw-cue encoding models do not encode cues relative to context, some predict lateral relationships at higher levels of processing. For example, Nearey's (1990) NAPP model of coarticulatory compensation uses context information to bias decisions at the level of diphones. In this model, raw cue-values are encoded and mapped onto phoneme categories, which are combined to form diphone representations. Biases to diphone responses are applied at this level based on information about coarticulatory context. Nearey argues against biases at the level of cue encoding, as they would make the models unnecessary complex — diphone biases are sufficient to account for the effects he describes.[4] Similarly, approaches using FLMP (Oden & Massaro, 1978) to handle context effects have suggested that context compensation could be performed at the level of categories via similar biases, though it could also be implemented at the level of cues in a way that is more consistent with other lateral models (Massaro & Oden, 1980).

Together, lateral models predict that only information at a particular stage of processing is used to compensate for contextual variability. For intrinsic and extrinsic models, this occurs at the level of cue encoding, and for raw-cue models it occurs at later stages.

2.2.2.2 Feedforward models

In contrast to lateral approaches, feedforward approaches (center row in Figure 2.1) suggest that context information is only combined with phonetic information via information flow from lower levels of processing to higher levels. As such, all approaches that posit feedforward explanations of context compensation are also raw-cue encoding models,

---

about gestures or via feedback from higher-level representations. Thus, along with the broader framework of direct realist approaches (Fowler, 1986) and the motor theory of speech perception (Liberman & Mattingly, 1985), these approaches are more difficult to classify using the scheme presented here, since they differ from previously discussed models in that the units of speech perception are articulatory or gestural from the earliest moments of processing — there is no level of acoustic cue encoding.

[4]Note that, because of this, NAPP differs from the intrinsic and extrinsic lateral models in that it encodes raw cues. Thus, a model classification scheme with more dimensions might be better able to classify NAPP relative to the other models. For simplification, it is categorized here based on the fact that it uses raw cue encoding and allows for lateral relationships at some stage of processing.

predicting that initial cue encoding is unaffected by contextual variability.

Exemplar models (Johnson, 1997; Goldinger, 1998) take contextual variability into account by integrating multiple cues via feedforward information to phonological feature representations. General cue-integration models, like FLMP (Oden & Massaro, 1978) can use a similar approach, though they can also handle contextual variability at the level of cue encoding or categories via processes more similar to the lateral interactions described for other models (Massaro & Oden, 1980). Similarly, C-CuRE, uses both raw-cue encoding via feedforward activation, as well as extrinsic encoding via feedback (see next section).

Recently, Toscano and McMurray (2010) presented a WGMM to explain rate compensation. In this model, phonetic categories are represented by Gaussian distributions along acoustic cue dimensions (e.g., VOT). The model learns the number of categories for a phonological contrast via an unsupervised, competitive learning process (McMurray, Tanenhaus, & Aslin, 2009) that discovers clusters of cue-values along a given dimension. Cues are weighted by their reliability (i.e., how distinct the categories along that dimension are), and weighted cues are used as input to a mixture containing information from multiple cues. This is used to create a continuous, but more abstract, representation of the phonological contrast supported by the cues, similar to the phonological features in FLMP. The posterior probability of the Gaussians in the combined mixture is used to estimate the likelihood of a response (e.g., /b/ vs. /p/) for a particular combination of cues.

Finally, in Smits' (2001b) HICAT model, raw-cues are encoded and mapped onto categories. Unlike some versions of FLMP, separate representations based on each cue are maintained until the likelihood of a particular category is computed. Also, unlike NAPP, HICAT does not include a category bias term at the level of diphones; the likelihood of responses for each segment are combined (via feedforward connections) to yield an overall estimate. Thus, contextual variability is handled by modulating feedforward information

from lower levels of processing using context information.[5]

To summarize, feedforward models all predict that information flows from lower levels of processing to higher ones. As a result, all of these are raw-cue encoding models, which predict that listeners do not encode cues relative to context.

2.2.2.3 Feedback models

Finally, some models propose that feedback from higher levels affects lower levels of processing to handle contextual variability (bottom row in Figure 2.1). Most models of extrinsic compensation fall into this category, though extrinsic models that rely on lateral interactions have also been proposed (see above).

Examples of this type of model include Gerstman (1968) and Lobanov's (1971) talker compensation models, which use the range and mean formant frequencies for a given talker, respectively, to recode formants into a normalized vowel space. Since information used for compensation is based on the properties of a particular talker's vowel space, these models suggest a feedback mechanism from more abstract information onto cue encoding (though these models were not described in terms of the direction of information flow in the system).

C-CuRE (McMurray & Jongman, 2011) presents another example of an extrinsic encoding model that uses feedback from category-level representations to compensate for contextual variation. C-CuRE initially encodes raw cue-values, which can be used by themselves via feedforward activation. In addition, information about context can be factored out of cue-level representations via feedback from knowledge about talkers and phonemes. For example, if a listener hears an ambiguous vowel between /i/ (high F2) and /ɪ/ (low F2) and can identify the talker, they can use this information to overcome the ambiguity. If the talker has higher F2 values on average, for example, they can then adjust their F2 estimate

---

[5]As with NAPP, the direction of information flow is difficult to classify in HICAT. Again, a more complex classification scheme would group these two models together given their differences from the other approaches, but, for simplification, it is discussed here as a feedforward, raw-cue encoding model.

do be lower than it actually is (compensating for that particular talker's higher F2 values), perceiving the previously-ambiguous sound as /ɪ/. Thus, expectation about the talker's voice allows listeners to handle contextual variability in their estimates of formant frequencies. Similar processes can be used to handle coarticulatory context: if a listener knows that the vowel was preceded by a particular consonant, they can factor out the expected contextual variability in the vowel that is due to that consonant.

This approach contrasts with other extrinsic encoding approaches (e.g., auditory contrast) that directly compare cue-values (via lateral processes). Instead, C-CuRE uses feedback based on expectations about cue-values. In addition, it differs from gestural parsing and auditory contrast approaches in that it does not make strong claims about the nature of the cue-level information (i.e., whether it is a gesture- or acoustic-based representation). Other types of models that use some form of feedback to update cue-level representations, such as normalized recurrence networks (Spivey, 2007; McMurray & Spivey, 1999) or iterative competitive learning networks (which use both raw and relativized cues as C-CuRE does Mozer, 1990), could be applied in similar ways.

While raw-cue encoding approaches could use feedback to account for context effects above the level of cue encoding, these have not generally been applied to the problem of contextual variability in speech. However, some models of spoken word recognition, like TRACE (McClelland & Elman, 1986), fit into this category. TRACE is an interactive activation network with acoustic cue, phoneme, and lexical levels of processing. The model encodes raw acoustic cue-values, but uses feedback from lexical representations on to phonemes to recognize words. This feedback could conceivably be useful to compensate for contextual variation at phoneme levels, though this has not been thoroughly examined.

Feedback models also suggest an interesting prediction about how different types of context information are handled. Recall that some types of contextual variability (e.g., talker identity) are relatively categorical and that others (e.g., speaking rate) are more continuous.

If contextual categories, such as whether a talker is male or female, are learned similarly to phonetic categories (i.e., via unsupervised learning), listeners may not have distinct representations at category levels for continuously-variable context information, like speaking rate. Clustering algorithms used to discover phonetic categories work well for multi-model cue distributions (McMurray, Aslin, & Toscano, 2009; Toscano & McMurray, 2010; Vallabha, McClelland, Pons, Werker, & Amano, 2007), such as those that provide cues to talker gender, and supervised learning approaches would be even more powerful at assigning abstract labels to particular constellations of phonetic cues associated with phonemes or talkers. However, unsupervised learning models would fail to form distinct categories for information from a unimodal or uniform distribution, such as those that indicate speaking rate. Thus, if listeners use feedback to compute relative cue-values, we might expect different effects for talker gender and speaking rate.[6] This is directly tested in Chapters 4 and 5.

Overall, this class of models predicts that listeners use feedback from later stages of processing to handle context effects. This could be done either at the level of cue encoding itself (suggesting relative encoding) or at an intermediate stage (suggesting raw-cue encoding). Thus, both relative and raw cue encoding approaches, by themselves, are consistent with feedback models.

2.3   Comparing proposed solutions

These three types of cue encoding (intrinsic, extrinsic, and raw-cues) and three directions of information flow (lateral, feedforward, and feedback) allow us to classify most models of contextual variation, as well as general approaches to speech perception. Different approaches make different predictions about the level of processing at which context compensation occurs and the nature of the information listeners use for compensation. More

---

[6]Listeners could also compute a more abstract representation of speaking rate that is not based on categories. Effects of this information could be considered a form of feedback, but they would not be logically distinct from lateral interactions in this classification scheme (since they reflect direct interactions between continuous dimensions).

generally, each approach differs in how those sources of information are combined. Thus, we can compare them using a general information processing framework that uses the same types of acoustic information but combines them in different ways for particular models.

Figure 2.1 depicts schematics of several possible models in a grid with type of encoding in the different columns and direction of information flow in the different rows. This illustrates the differences in between the general approaches for handling contextual variability and notes where some specific models fall under this classification system. Cells that are not possible (e.g., intrinsic encoding via feedforward activation) are listed as "n/a". Note that some models may fall under multiple categories but are listed in the category that most clearly emphasizes the key features of that model.

The top-left corner of the figure corresponds to lateral, intrinsic encoding models. This includes the approaches that have been termed "invariance" (Stevens & Blumstein, 1978; Blumstein & Stevens, 1979) or "compound-cue" approaches. The top-center panel depicts lateral, extrinsic encoding models, such as auditory contrast approaches (Lotto & Kluender, 1998). The top-right panel illustrates lateral, raw-cue models, like NAPP (Nearey, 1990). The middle row shows feedforward approaches, all of which use raw-cue encoding. This includes WGMM, HICAT, and some implementations of FLMP. The bottom row depicts the two feedback approaches, including those that use extrinsic encoding (bottom-center), like C-CuRE (McMurray & Jongman, 2011), and those that use raw cues (bottom-right), like TRACE (McClelland & Elman, 1986).

2.3.1 Predictions

Given this classification system, we can distinguish between most of the proposed models on the basis of these two principles by examining how cues are encoded. First, the three encoding approaches make different predictions about when during speech processing we should see effects of context information. If listeners use either intrinsic or extrinsic encoding, contextual variability should be handled at the level of cue encoding itself, such

that cues are encoded relative to preceding context. In contrast, raw-cue models predict that acoustic cues are encoded veridically and context information is only taken into account at later stages of processing. Thus, if listeners use this approach, cue encoding should not be affected by context. By studying cue encoding in different contexts we can distinguish between the relative (intrinsic/extrinsic) and absolute (raw-cue) approaches.

Second, the three directions of information flow may make different predictions for different types of context information. If preceding contexts provides listeners with information about a particular category (e.g., the gender of the talker), feedback approaches predict that this information can influence lower levels of processing, including cue encoding. However, if context instead provides more continuous information (e.g., variation in speaking rate), feedback from categories may not affect cue encoding. Continuous information could be used by lateral interaction and feedforward approaches, but feedback models may not show an effect of cue encoding in this case. Thus, by examining different types of context information (e.g., talker vs. speaking rate variability), we can distinguish between feedback and lateral/feedforward approaches.

Assessing these differences requires us to look at cue encoding and categorization during speech processing for different types of context effects. The next chapter presents an ERP approach designed to measure processing at these two levels and presents initial experiments examining responses to different speech sounds to determine whether we will be able to use them to measure cue encoding in subsequent experiments.

CHAPTER 3
EVENT-RELATED POTENTIALS TO SPEECH SOUNDS

As stated in the introduction, the main goal of this dissertation is to simultaneously examine effects of various types of contextual variability (speaking rate, talker, coarticulation) on cue-level and category-level responses. While many of the methodological details of each experiment will be specific to these individual factors, the overall logic of the experiments is quite similar, and the use of the ERP technique is largely the same. The approach used here is similar to that used by Toscano et al. (2010) to examine whether VOT encoding and categorization are sensitive to continuous acoustic differences: using the auditory N1 as a measure of cue encoding and the P3 as a measure of categorization and post-perceptual processing. To do this, several experiments will examine acoustic variation along different cue dimensions, specifically VOT and formant dimensions, in the context of particular sources of context information (talker, rate, or coarticulatory context).

The distinction between ERP responses that reflect cue encoding and those that reflect categorization is similar to the distinction made by Picton and Hillyard (1974) for ERP responses to nonspeech auditory stimuli. They presented listeners with click trains that contained an occasional click that was either at a lower intensity (1-5 dB less) or was omitted entirely. Participants were instructed to either detect and count the infrequent clicks or to ignore the auditory stimuli. They found larger N1 and P3 components in the attended than in the unattended condition, suggesting that attention operates as early as the time of the N1. In addition, a smaller N1 was observed for the infrequent, lower intensity clicks relative to the more frequent, louder ones, and no N1 was observed when the stimulus was omitted. In contrast, the P3 was similar in both the lower-intensity and omitted conditions.

From these results, Picton and Hillyard suggested that the N1 and P3 may correspond to two different stages of processing: (1) an evoked sensory response during which the auditory input can be compared to expectations (the N1), and (2) a decision process that

is not dependent on the particular acoustic properties of the input but is affected by the properties of the task (the P3). This corresponds very closely to the proposal made by Toscano et al. (2010) for speech sounds. They found that the N1 was dependent on the VOT of the stimuli and did not vary as a function of the task-defined target words. Larger N1s were observed for short VOTs and smaller N1s for long VOTs, regardless of whether listeners were monitoring for a voiced or voiceless word. In contrast, the size of the P3 was determined by the distance of a given VOT from the task-defined target word. That is, if *peach* (which is consistent with a VOT of 40 ms) was the target, P3 was largest for the 40 ms stimuli. In contrast if *beach* was the target, P3 was largest for the 0 ms stimuli.

These results map onto Picton and Hillyard's (1974) indexing functions quite well, and this suggests that we may be able to use the N1/P3 paradigm to measure cue encoding and categorization for other speech stimuli also. However, there are several additional methodological points that must be addressed to confirm that this is the case. Thus, before presenting each of the sets of experiments examining different types of context effects, I will first discuss several points here that relate to the overall study and present two experiments designed to examine ERP responses to VOT and formant differences, which will be examined in subsequent experiments.

First, I present a follow-up to the analyses of Toscano et al. (2010) to provide additional evidence demonstrating that the P3 component can be used as a measure of phonological categorization. Second, I discuss the issue of overlapping ERP components that are caused by temporally adjacent stimuli. This issue is relevant to the current study because conclusions about context effects depend on responses to a stimulus as a function of a preceding stimulus.

Finally, I discuss the phonetic contrasts that will be used in the experiments and whether differences along those acoustic dimensions produce differences in the auditory N1. Two experiments, one examining vowel differences and one looking at voicing contrasts, are

presented to determine whether differences in the N1 can be seen for these stimuli.

3.1 Phonological categorization and P3 amplitude

In addition to measuring cue encoding, we would also like to measure category-level processing. As discussed in Chapter 1, this is the level that most behavioral measures of context compensation (e.g., phoneme identification) have assessed. Thus, measuring this using ERPs seems to serve primarily as a an additional way to measure these effects. At the same time, there has been little electrophysiological work examining this level of phonetic categorization, so an ERP approach may offer an important extension of these results.

One of the original goals of Toscano et al. (2010) was to use the P3 component to measure categorization of speech stimuli. This was done using a target detection task in which listeners categorize whether a given stimulus matched a target category. Targets occur infrequently (≈25% of the time), producing a P3 component on target trials. The results from Toscano et al. suggest that variation in P3 amplitude reflects phonological category differences (rather than acoustic differences) — P3 amplitude varied with overall distance from the target endpoint for the relevant stimulus continuum (e.g., the *beach-peach* continuum if *beach* was the target) rather than with VOT itself. In addition, when the data were recoded to account for differences in individual participant's category boundaries, an effect of VOT within each voicing category was found.

To further establish that variation in P3 amplitude reflects differences in listeners' phonological categories, an additional analysis was run on the data from Toscano et al. (2010). If P3 amplitude reflects category information, distance from the target endpoint relative to participants' individual category boundaries, should produce the greatest variation in the size of the P3. In contrast, if the P3 reflects some other difference between the stimuli, a larger effect would be observed when the data are divided at a different VOT step than the participant's category boundary.

To test whether this was the case, the data were recoded by shifting participant's

true category boundaries up to two VOT steps in either direction. This produced five data sets from the original data: those grouped according to participants' category boundaries (relative VOT, or rVOT), those grouped based on their boundaries shifted one step toward the voiced end of the continua (rVOT-1), two steps toward the voiced end (rVOT-2), one step toward the voiceless end (rVOT+1), and two steps toward the voiceless end (rVOT+2).

These five data sets were analyzed with linear mixed-effects models using the lme4 package (Bates & Sarkar, 2011) in R (R Development Core Team, 2011). Mixed-effects models provide a more appropriate statistical analysis than ANOVA for these data, since they allow continuous variables (like rVOT) to be coded correctly (see Jaeger, 2008). In this and subsequent mixed-effects models used in this dissertation, participant was entered as a random effect (on the intercept) and continuous variables were centered, but not normalized. This means that the reported statistics are unstandardized (b) rather than standardized (β) coefficients, so the value indicates differences in the dependent variable as a function of each level of the independent variable (e.g., coefficients for P3 amplitudes, presented here, reflect the size of the difference in μV per rVOT step).

In this model, participant was entered as a random effect and the absolute value of rVOT was entered as a fixed effect (the data were collapsed across the two VOT continua used in the experiment).[1] The results showed a significant effect for the data based on participant's true VOT boundaries (b = 0.167, p_{MCMC} = 0.007) and for the rVOT+1 data set (b = 0.138, p_{MCMC} = 0.01). Effects for the other data sets were non-significant. In addition, the log likelihood of the rVOT model was higher than that of the other models (rVOT: -357.5; rVOT-1: -361.7; rVOT-2: -363.0; rVOT+1: -358.2; rVOT+2: -361.4).

These results suggest that participants' category boundaries produced the largest variation in P3 amplitude, providing additional evidence that the P3 serves as an index of categorization in this task. Given this, a target detection task will be used to assess whether

---

[1]Note that, unlike the analyses from Toscano et al., these data were not grouped by response.

listeners' categorization of other speech sounds, specifically formant frequency differences (Experiments 5, 6 and 8) are graded in the same way categorization of VOT is.

3.2   Separating overlapping ERP components

A second general methodological issue is the problem of overlapping ERP components. Unlike Toscano et al. (2010), the contextual variability experiments will examine the influence of two acoustic events (e.g., the preceding sentence and a target word). This raises the possibility that ERP responses to the target stimulus could be due to overlap from components related to processing the preceding context, rather than context compensation effects. Because ERP components can occur several hundred milliseconds after a stimulus, it is important to know that any differences in components to the target stimulus are not actually due to overlapping components from the different contexts. This is a particularly relevant issue for examining effects of local context (i.e., coarticulatory context), since the onset of both stimuli occurs relatively close in time. For example, to examine coarticulatory effects, we can look at differences in N1 amplitude to the onset of the vocalic portion of a fricative-vowel syllable as a function of the coarticulatory information in the frication. A difference in the N1 to the vowel onset could be due to context-relative cue encoding (the predicted effect) or due to the acoustic differences between the two frication segments.

To rule out this possibility, the Adjar procedure (Woldorff, 1993; Hopfinger & Mangun, 1998) will be used to separate ERP components to the context from components to the target if overlap is observed (i.e., a non-zero baseline before the onset of the target stimulus). Adjar involves iteratively estimating and removing overlap between components caused by two temporally-adjacent events (e.g., two stimuli). This requires that the ISI between the two stimuli (the sentence onset and target word onset) varies over some interval. This ISI jitter by itself helps to reduce overlap by smearing out the components caused by one stimulus in the ERP waveform to the other stimulus. However, if the stimuli are presented close together, there can still be overlap between the components. This is particularly problematic

for the second stimulus, as components from the first can disrupt the baseline voltage.

To use Adjar to remove component overlap, the average ERP waveform time-locked to the onset of the second stimulus (S2; e.g., the onset of the target word) is obtained. This is used as the current best estimate of the ERP components due to S2. Next, this waveform is shifted forward in time at each of the ISI steps. This produces an estimate of the overlap from S2 on the first stimulus (S1; e.g., the onset of the carrier sentence). This average is then subtracted from the ERP to S1 to remove the overlap. The same procedure is used to remove overlap from S1 on the S2 ERP, shifting the S1 waveform backward in time at each of the ISIs and subtracting it from the original S2 ERP. This procedure is repeated until the change in the overlap estimate is near zero (here, if the maximum absolute difference is $< 1^{-4}$).

3.3 Auditory N1 response to speech sounds

Finally, in order to examine effects on cue encoding, we must use stimuli that vary along acoustic dimensions that produce observable differences in the N1. A wide range of auditory stimuli generate a fronto-central N1 (Näätanen, 1987). The N1 varies as a function of stimulus frequency (Antinoro & Skinner, 1968; Picton, Woods, & Proulx, 1978), intensity (H. Davis & Zerlin, 1966; Picton et al., 1978), rate of stimulation for repeating stimuli (Butler, 1973; Picton et al., 1978), phonetic differences in speech sounds (Wood, 1971; Lawson & Gaillard, 1981; Sharma & Dorman, 1999), and the signal-to-noise ratio of speech stimuli (B. A. Martin, Sigal, Kurtzberg, & Stapells, 1997).

In addition, there are several non-auditory factors that affect the N1. As discussed above, the N1 is affected by selective attention (Hansen & Hillyard, 1980; Öhman & Lader, 1972; Picton & Hillyard, 1974), with larger N1s observed when participants are attending to a stimulus than when they are not. Similarly, the N1 is affected by non-auditory but stimulus-driven factors, such as the visual stimulus in audiovisual speech perception (Pilling, 2009; van Wassenhove, Grant, & Poeppel, 2005; Besle, Fort, Delpuech, & Giard, 2004). Visual information conveys cues that listeners use during speech perception (McGurk & MacDonald,

1976), and van Wassenhove et al. (2005) examined whether visual speech information has an effect on the auditory N1 by presenting listeners with audio-only, visual-only, and audiovisual speech samples. They found that the N1 response to a given syllable was smaller in amplitude and earlier in the audiovisual condition than in the audio-only condition; no N1 response was observed in the visual-only condition. This suggests that non-auditory information can also affect the N1.

Another issue pertinent to the present study is the effect of specific acoustic differences on the N1. Picton et al. (1978) observed changes in N1 amplitude for tones varying in frequency. Specifically, they found larger N1s are produced by lower-frequency stimuli. They suggested that this frequency effect could be due to the organization of the peripheral auditory system — the frequency mapping of the cochlea is warped such that lower frequencies are disproportionately represented. They also suggested that the cortical neurons that respond to higher frequencies may be oriented in such a way that their electric fields are not easily observable at the scalp. Studies using positron emission tomography to localize activity in auditory cortex (Lauter, Herscovitch, Formby, & Raichle, 1985), as well as ERP localization studies (Bertrand, Perrin, & Pernier, 1991), support this latter view — neural populations the respond to high-frequency sounds are located posterior and deeper than those that respond to low-frequency sounds. Thus, responses high-frequency sounds would be more difficult to observe at the scalp due to the orientation of their generators.

These results suggest that the N1 to high frequency speech stimuli might also be small. Moreover, it suggests that the differences in stimuli along a continuum varying in a high-frequency cue may be difficult to assess. Tremblay, Friesen, Martin, and Wright (2003) examined ERPs in response to an /s/-/ʃ/ ("s" vs. "sh") distinction and did not find any differences in the N1 (though they did find a difference for this contrast at a later time point). Similarly, our own pilot work examining ERP responses to /s/-/ʃ/ stimuli also showed no differences in the N1. Thus, it may be that not all differences in phonetic cues produce

differences in N1 amplitude, even if listeners can perceive the difference between the sounds.

Several ERP experiments examining differences in the N1 to speech sounds have focused on VOT (Sharma & Dorman, 1999, 2000; Sharma, Marsh, & Dorman, 2000; Frye et al., 2007; Steinschneider, Volkov, Noh, Garell, & Howard, 1999; Toscano et al., 2010). VOT is typically defined as a temporal cue (i.e., the time between two events), and indeed some studies show two N1 peaks for long VOTs, with the first N1 corresponding to the release burst and the second N1 to the onset of voicing. However, VOT may also be described in terms of spectral information. Frequency differences between sounds with short and long VOTs can be seen by looking at the short-term average spectrum over the beginning of the syllable. This spectral measure of VOT shows large differences in low-frequency information, since the short-VOT stimuli contain more voicing energy. Indeed, this is a likely explanation for the linear effect of VOT and lack of a double-peaked response in Frye et al. (2007) and Toscano et al. (2010).

Given this, it seems logical to examine variation in VOT in the present set of experiments. VOT provides a useful cue for examining the effects of speaking rate: it is primarily a temporal cue, it is profoundly affected by the rate of the surrounding speech, and listeners must compensate for speaking rate variability when interpreting VOT (Miller & Grosjean, 1981; Miller & Dexter, 1988; McMurray, Clayards, Tanenhaus, & Aslin, 2008; Toscano & Mc-Murray, 2011a). In addition, VOT values also vary between male and female talkers, such that women have longer VOTs than men (Swartz, 1992; Ryalls, Zipprer, & Baldauff, 1997; Whiteside & Irving, 1998). Thus, talker context could also produce effects on VOT encoding.

ERP responses to other phonetic cues, such as formant differences for vowels (Tremblay et al., 2003), have also been found. Formants vary significantly between talkers, and talker-specific factors can account for up to 80% of the variance in formant frequency (Cole et al., 2010). In addition, formant frequencies for vowels are also dependent on coarticulatory context. Therefore, vowel differences may provide a useful phonological contrast for examining

talker context effects as well.

Given these results, vowel and voicing distinctions allow us to examine several types of context effects. Both VOT and formant frequency are affected by talker gender, and VOT is affected by rate. This allows us to test alternative explanations for observed context effects on the N1, since, if listeners do compensate for contextual variability during cue processing, we would expect an effect on both types of cues for talker differences, but only an effect on VOT for rate differences.

In the next two sections, two experiments looking at differences in N1 responses to stimuli varying in VOT (i.e., a voicing contrast) and formant frequency (a vowel contrast) are presented. Experiment 1 examines differences between the vowels /æ/ and /ɛ/, and Experiment 2 examines differences between the consonants /b/ and /p/.

3.4 Experiment 1: N1 response to vowel differences

The purpose of this experiment is to establish that /æ/-/ɛ/ vowel differences produce differences in N1 amplitude. This will allow us to establish that this vowel distinction can be used for subsequent ERP experiments. To test this, ERPs were recorded while participants listened to natural recordings of the words *axe*, *ex*, *had*, and *head* made by one male and one female talker (to see whether talker information interacted with listeners' N1 responses to the two vowels).

3.4.1 Methods

3.4.1.1 *Participants*

Nine people participated in the experiment. Participants were recruited from the University of Iowa community according to University human subjects protocols, provided informed consent, and were either compensated $15 per hour or received course credit for participation. Participants reported English as their only native language, normal hearing, and normal or corrected-to-normal vision.

3.4.1.2   Design

Participants performed a 2AFC word identification task. Note that this task should produce an auditory N1, but not a P3 component. Stimuli contained two vowels (/ɛ/ and /æ/) and two word pairs (*ex-axe* and *head-had*), each of which was produced by two talkers (one female, one male), for a total of eight stimuli. Each word pair was presented in different halves of the experiment, and the order of presentation was alternated between participants. Within each half, stimuli were presented in random order. Each stimulus was repeated 45 times for a total of 360 trials. The experiment lasted approximately one hour and was run in a single session with another experiment in our lab (which also lasted one hour).

3.4.1.3   Stimuli

Two talkers recorded the stimuli in a sound-attenuated room with a Kay CSL 4501. Recordings were made using Praat (Boersma & Weenink, 2010) and were sampled at 44.1 kHz. Each talker recorded several tokens of the four words, and the token with the best audio quality was selected for the experiment.

Within each word pair and talker gender condition, /æ/ and /ɛ/ vowels were equated for length to ensure that differences in ERP component latency were not due to differences in the length of the stimuli. This was done by first measuring the length of the vowel in each token. Next, for the *ex-axe* word pair, the longer token from each talker was shortened so that it had the same duration as the shortest token from that talker (*ex* was shorter than *axe* in both cases). This was done by removing periodic energy from the onset of the longer stimulus, cutting the sound at the zero-crossing closest to the desired time point. The same procedure was applied to the *head-had* stimuli (*head* was shorter than *had* for both talkers).

The /h/ from the *head* token was also replaced by the /h/ from *had* to ensure that the stimuli had the same /h/ for each talker. Similarly, the consonant following the vowel from the /ɛ/-tokens replaced the consonant from the /æ/-tokens. Stimuli were also normalized for intensity to ensure that ERP differences would not be due to differences in sound level.

3.4.1.4 Procedure

Participants were seated in an electrically-shielded, sound-attenuated booth. Stimuli were presented over Sennheiser HD 555 headphones. This model of headphones was used in all subsequent experiments also. Participants indicated their response using two buttons mounted on a board placed in front of them that were connected to the stimulus presentation computer. An LCD projector outside the booth was used to present instructions and visual stimuli to the participant on a screen inside the booth. Stimulus presentation and recording of participants' behavioral responses was handled using the Psychophysics Toolbox (Brainard, 1997) and custom MATLAB scripts.

At the beginning of each trial, a fixation point and the relevant word pair ("axe" and "ex" or "had" and "head") appeared on the screen. The words were located to the left and right of the fixation point and indicated which button corresponded to which sound (i.e., if "axe" was on the left side of the screen, the left button was used to make an /æ/ response). The response-button mapping was held constant throughout the experiment and was alternated between participants. 750 ms after the fixation point and letters appeared, the auditory stimulus was presented. Participants then indicated their response by pressing one of the buttons, and the screen was cleared. If the participant waited more than 3000 ms after stimulus onset to respond, they heard an error buzzer and the trial ended (this occurred very infrequently).

To minimize ocular artifacts, participants were instructed to maintain their gaze on the fixation point while it was on the screen. The inter-trial interval varied randomly between 1250 and 1750 ms, allowing participants sufficient time to blink between trials. Trials were divided into blocks of 15 trials each, and participants were given the opportunity to take a break between each block. Halfway through the experiment, the participants took a longer break.

3.4.1.5   EEG recording

ERPs were recorded from 11 electrode sites (International 10-20 System sites F3, F4, Fz, Cz, P3, Pz, P4, PO3, and PO4; and two sites approximately 1 cm anterior and superior to C3 and C4). EEG channels were referenced to the left mastoid during recording and re-referenced to the average of the two mastoids after recording. Horizontal electrooculogram (EOG) recordings were made using two electrodes located approximately 1 cm lateral to the external canthus of each eye. Vertical EOG recordings were made using an electrode located at FP2. Impedance was 5 kΩ or less at all sites. Recordings were made with an SA Instrumentation bioamplifer with a low-pass filter at 100 Hz for all channels and a high-pass filter at 0.01 Hz for the EEG channels, 0.3 Hz for the vertical EOG channel, and 0.1 Hz for the horizontal EOG channels. Data was digitized at a sampling rate of 250 Hz and saved to disk, along with information about stimulus conditions, event times, and responses, using a custom-made program.

3.4.1.6   Data processing

Data were analyzed using the ERPLAB plugin (Luck & Lopez-Calderon, 2010) to the EEGLAB toolbox (Delorme & Makeig, 2004) for MATLAB. Trials containing ocular artifacts were rejected if the peak-to-peak voltage between -100 and 400 ms exceeded 74 μV for either of the EOG channels or 150 μV for any of the EEG channels. Approximately 5% of the trials were rejected for a given subject (similar artifact rejection levels were observed in subsequent experiments also). The baseline for each epoch was the average voltage 100 ms before the onset of the auditory stimulus.

3.4.2   Results

Listeners correctly categorized the vowel endpoints for both talkers and word pairs (mean accuracy: 96.3%).

The acoustic differences between the two word pairs (i.e., *ex/axe* vs. *head/had*)

are not directly comparable, since one has an /h/ at onset and the other does not and the vowel differences between the two may not be the same magnitude or may be driven by different cues. Because of this, the two word pairs were analyzed separately. Figure 3.1 shows grandaverage ERP waveforms for each word pair and vowel collapsed across talker. Overall, the /æ/ stimuli appear to have larger N1 amplitudes than the /ɛ/ stimuli.

Mean N1 amplitude was computed as the average voltage across the three frontal channels (F3, Fz, and F4) from 75 to 125 ms after the onset of periodic voicing (i.e., the onset of the vocalic portion). This time range corresponded to 75-125 ms after stimulus onset for the *ex-axe* stimuli and 223-273 ms after stimulus onset for the *head-had* stimuli (due to the word-initial /h/). Figure 3.2 shows mean N1 amplitude as a function of talker gender, word pair, and vowel.

For the *ex/axe* word pair, mean amplitude was -1.37 μV for the *axe* stimuli and -1.02 μV for the *ex* stimuli. A 2 (talker) x 2 (endpoint) within-subjects ANOVA showed that this difference was significant (F(1,8)=7.38, p=0.026). Neither the effect of talker gender (F<1) nor the interaction (F<1) were significant.

For the *head-had* stimuli, mean amplitude was slightly smaller for *had* (-0.04 μV) than for *head* (0.20 μV). This is the same pattern observed for the *ex/axe* stimuli, but there is clearly considerable overlap from the preceding /h/, as evidenced by the lack of a flat waveform preceding the onset of the vocalic portion. A 2 (talker) x 2 (endpoint) within-subjects ANOVA did not find a significant effect of vowel (F(1,7)=1.24, p=0.302), but there was a marginal effect of talker gender (F(1,7)=4.71, p=0.067) with larger N1 responses for the male talker than the female talker. The direction of this effect is the same as that seen for VOT stimuli (Toscano et al., 2010) and tones (Picton et al., 1978), with lower-frequency sounds (i.e., the male talker, short VOTs, and low frequency tones) producing larger N1s. The interaction was not significant (F<1).[2]

---

[2]One participant was missing data for the *had-head* stimuli due to equipment problems and was excluded from this analysis.
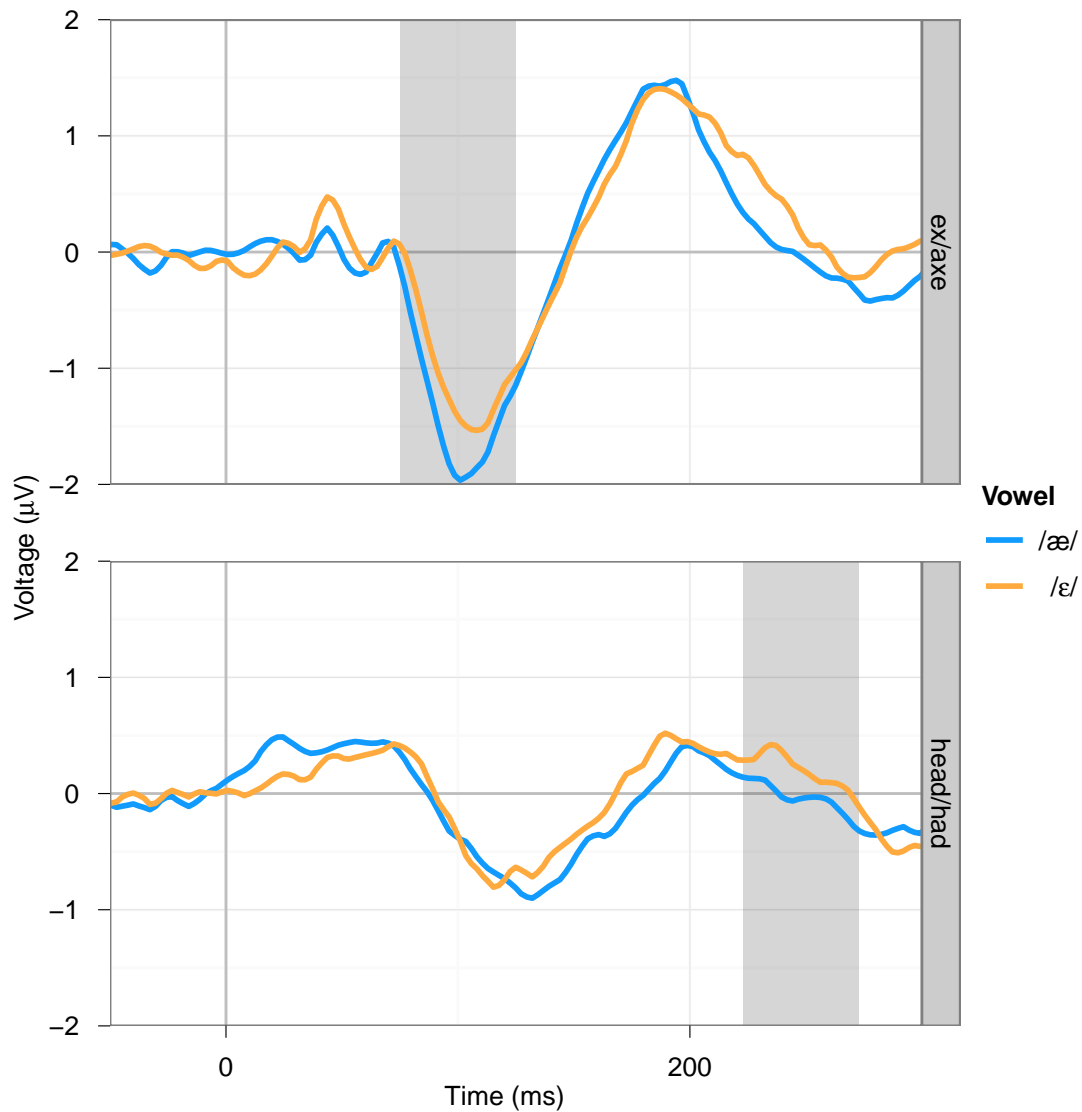
Figure 3.1: Experiment 1 results — ERP waveforms. Grandaverage waveforms for the average of the three frontal channels as a function of word pair (*ex/axe* or *head/had*) and vowel (/ɛ/ or /æ/). Waveforms are time-locked to the onset of the word. Shaded areas indicate the time range used to compute mean N1 amplitude for the vowel differences. This is later for the *head/had* stimuli due to the word-initial /h/. Positive is plotted up; this is true of other ERP figures as well. This figure, as well as subsequent figures showing experiment results, were generated using the ggplot2 package in R (Wickham, 2009).
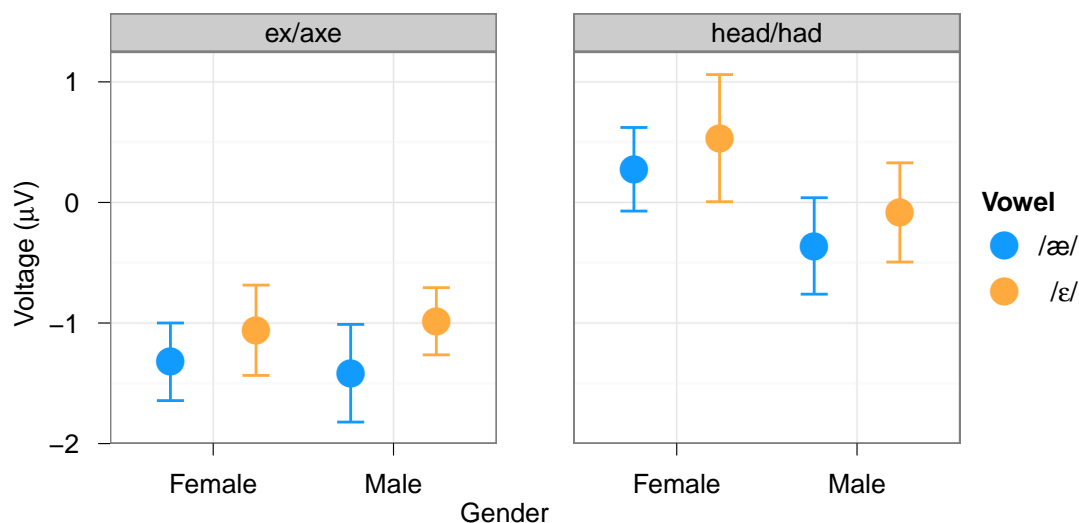
Figure 3.2: Experiment 1 results — N1 amplitude. Mean amplitude as a function of word pair, talker, and vowel. The N1 is larger for the *axe* stimuli and the *ex* stimuli. Shaded areas for each condition indicate standard error.

3.4.3 Discussion

These results demonstrate that differences in the N1 can be seen for the /ɛ/-/æ/ distinction. Experiment 3 will examine this effect more closely to determine whether it reflects continuous variation in acoustic information, independently of listener's phonological categories (as it does for VOT).

The results also suggest that variation in N1 amplitude may not correspond simply to overall frequency differences, such that lower-frequency sounds produce larger N1s. Recall that in Toscano et al. (2010) shorter VOTs (i.e., lower-frequency sounds) produced larger N1s, and that similar results have been found for tones (Picton et al., 1978). However, in the present experiment, this was not the case: the overall frequency difference between the *axe* and *ex* stimuli was estimated by computing the center of gravity for each sound, and the *axe* stimuli (which produced larger N1s) had a higher spectral mean than the *ex* stimuli. These differences are mostly driven by the stimuli from the male talker who had a mean

frequency of 1777 Hz for *ex* and 2205 Hz for *axe*. The female talker showed little difference in mean frequency between the two vowels, but had a slightly lower mean frequency for *axe* (1137 Hz) than for *ex* (1184 Hz). Thus, differences in N1 amplitude may reflect variation in acoustic information within particular frequency bands or to variation along particular cue dimensions, rather than simply varying with the overall frequency of the sound.

Regardless of the direction of the effects though, the results of the present experiment demonstrate that vowel sounds have an effect on N1 amplitude. This suggests that they may be a viable candidate for examining cue encoding and context compensation in subsequent experiments.

3.5 Experiment 2: N1 responses to VOT differences

The second acoustic cue dimension that will be used to examine context effects is VOT, one of the main cues to voicing distinctions in English and a cue that is affected by both talker gender (Swartz, 1992) and speaking rate differences (Allen & Miller, 1999). The results of Toscano et al. (2010) demonstrated that, for synthetic speech, differences in VOT produce corresponding changes in N1 amplitude. Most previous work has also examined effects of VOT on the N1, finding effects for both negative VOT (pre-voiced) differences (Sharma & Dorman, 2000) and positive VOT differences (Sharma & Dorman, 1999; Tremblay et al., 2003; Frye et al., 2007).

However, several researchers have found differences in N1 morphology for short and long VOTs, in which short VOTs produce a single-peaked N1 and long VOTs produce a double-peaked N1 (Sharma & Dorman, 1999; Sharma et al., 2000; Steinschneider et al., 1999), and have suggested that this reflects auditory discontinuities in the processing of VOT. One potential issue is that the range of VOTs used in these studies was much longer (0 to 80 ms) than the range of VOTs that reliably distinguishes voicing categories for bilabial stops in English (0 to 40 ms). These studies found that the change in N1 morphology occurred between 40 and 50 ms VOT, well beyond the typical voicing category boundary. Moreover,

Sharma et al. (2000) observed that the presence of a single- or double-peaked N1 did not predict listeners' boundary between voiced and voiceless sounds. Finally, despite noting changes in N1 morphology, Sharma and Dorman (1999) and Sharma et al. (2000) also found that latency of the latest N1 peak was linearly correlated with VOT.

In contrast to these results, Toscano et al. (2010) found a single-peaked N1 across their entire VOT continuum. This may have been because (1) they varied VOTs from 0 to 40 ms, and (2) they used stimuli with low-amplitude bursts. The two N1 peaks seen in previous studies could be caused by two sudden changes in the acoustic properties of the stimuli, with the first peak corresponding to the release burst and the second peak to the onset of voicing. Because these two events occur closer together at short VOTs, only a single N1 peak may be observed in the ERP waveform. In addition, by using a low-amplitude burst, Toscano et al. (2010) may have eliminated or minimized the N1 to that event.

Because naturally-produced stimuli are used in these experiments, it is more difficult to manipulate the amplitude of the burst. Thus, it is important to determine whether a single-peaked N1 is observed for these stimuli. The goals of this experiment are (1) to see if long-VOT stimuli produce a single- or double-peaked response, and (2) if a single-peak is observed, to see if there are differences in mean N1 amplitude between long-VOT and short-VOT conditions.

3.5.1 Methods

3.5.1.1 *Participants*

Six people participated in the experiment. Participants were recruited using the same procedures used in Experiment 1, met the same criteria, provided informed consent, and received course credit or $15 per hour for participating.

3.5.1.2   Design

Participants performed a 2AFC word phoneme task in which they listened to spoken words and indicated whether the word started with "b" or "p" while ERPs were recorded. Stimuli consisted of six sets of /b/-/p/ minimal pair words (*bath-path*, *beach-peach*, *beak-peak*, *bet-pet*, *bike-pike*, and *buck-puck*). Each stimulus was presented 45 times in random order for a total of 540 trials. The experiment took approximately 60 minutes and was completed in a single session.

3.5.1.3   Stimuli

Stimuli were recorded by a male talker in a sound-attenuated room using a Marantz PMD670 recorder. Recordings were made at a sampling rate of 22.05 kHz and saved to a computer. Several tokens of each word were recorded, and the tokens with the best audio quality were selected for the experiment. Sound files were edited using Praat (Boersma & Weenink, 2010).

VOT continua were created from these recordings for another experiment; only the endpoint VOTs (approximately 0 and 40 ms) were used here. Continua were created by cross-splicing the voiced and voiceless tokens (McMurray, Aslin, Tanenhaus, Spivey, & Subik, 2008). Each token was marked at zero-crossings in approximately 5 ms steps from the onset of the word to 40 ms after onset, and nine-step VOT continua were created. To construct a given step in the VOT continua, the appropriate amount of (voiced) material we excised from the beginning of each voiced token and replaced with a similarly long (aspirated) portion from the voiceless.

3.5.1.4   Procedure, EEG recording, and data processing

The experimental setup, EEG recording, and data processing procedures were the same as in Experiment 1. The task was also the same, except that participants identified the first letter of the word they heard as either "b" or "p".
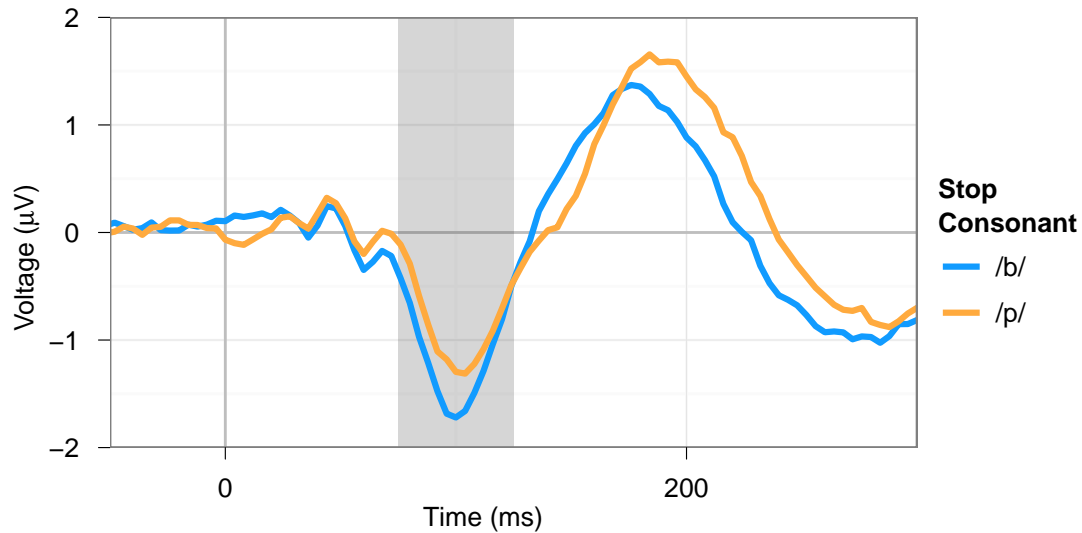
Figure 3.3: Experiment 2 results — ERP waveforms. Grandaverage waveforms for the /b/ (≈0 ms VOT) and /p/ (≈40 ms VOT) stimuli, averaged across the six word pairs. N1 amplitude is greater for the /b/ stimuli, and a single N1 peak is observed for both sets of stimuli. Raw averages are plotted in this figure to illustrate the absence of a double-peaked N1.

3.5.2 Results

Participants accurately identified the stimuli as starting with a /b/ or /p/ for each word pair (mean accuracy: 99.3%).

Figure 3.3 shows grandaverage ERP waveforms for the two VOT conditions. A single N1 peak can be seen for both the short and long VOTs. Mean N1 amplitude was calculated as the mean voltage across the frontal channels from 75 to 125 ms post-stimulus. Figure 3.4 shows mean N1 amplitude for each word pair and VOT endpoint.

Mean N1 amplitude was greater for the voiced condition (-0.84 μV) than for the voiceless condition (-0.65 μV). A 2 (VOT step) x 6 (word pair) repeated-measures ANOVA confirmed this, showing a significant effect of voicing endpoint ($F(1,5)=10.47$, $p=0.023$). This supports previous results showing that N1 amplitude is sensitive to VOT differences. Neither the effect of word pair ($F(5,25)=1.65$, $p=0.185$) nor the interaction ($F<1$) were significant.
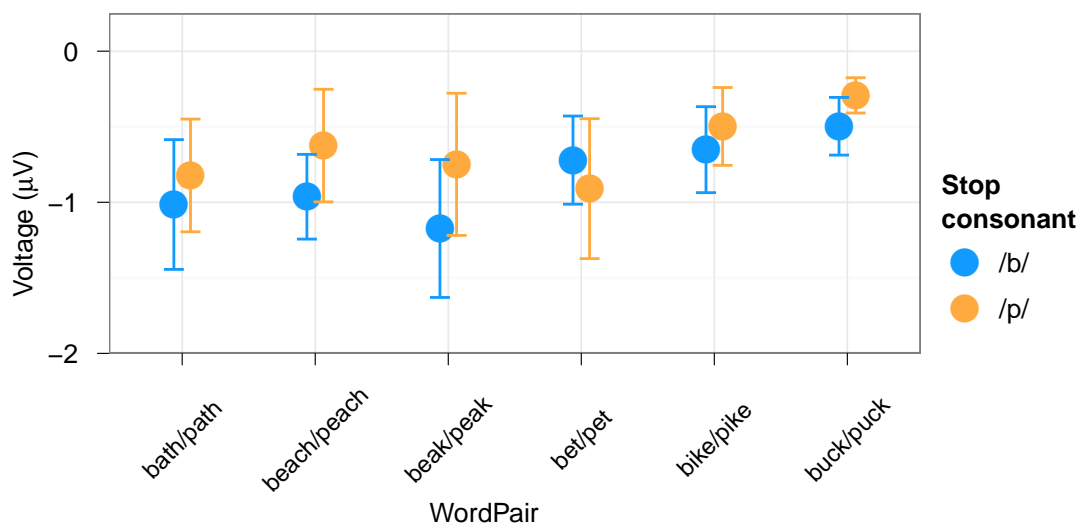
Figure 3.4: Experiment 2 results — N1 amplitude. Mean amplitudes for each word pair as a function of voicing condition (/b/ vs. /p/).

3.5.3 Discussion

The results of this experiment demonstrate that VOT differences in natural speech produce changes in N1 amplitude in the same direction seen with synthetic speech (larger N1s for short VOTs). In addition, a single-peaked N1 is observed for both endpoints, suggesting that the longer VOTs are not so long that they produce a double-peaked response and that the amplitude of the bursts in the natural stimuli is small enough that an N1 to the release burst does not obscure effects of VOT differences. This extends the results of Toscano et al. (2010) showing that similar effects can be observed for natural speech.

Overall, these results suggest that we can observe differences in N1 amplitude as a function of VOT for natural stimuli and demonstrates that they can be used in the experiments examining talker and rate compensation.

3.6 General discussion

The findings presented in this chapter provide the groundwork for examining the effects of context on cue encoding and categorization during speech perception. The analyses of the P3 data from Toscano et al. (2010) show that variation in P3 amplitude reflects differences in VOT relative to listeners' individual phonological categories. This provides additional support for the idea that the P3 can be used as measure of phonological categorization in a target detection task.

In addition, the results of Experiments 1 and 2 provide evidence that the acoustic differences of interest — formant frequency and VOT differences — produce measurable changes in the auditory N1. Given these results, we can use these cues to examine different types of context information on cue encoding.

Together, the P3 and N1 responses in the target detection task, along with the use of the Adjar procedure to remove component overlap, provide a set of tools that can allow us to examine context effects at the level of cues and categories. The next chapter presents a series of experiments using this approach to look at listeners' use of VOT and F1 as cues to voicing and vowel contrasts as a function of talker gender context.

CHAPTER 4
TALKER IDENTITY

4.1   Background

The next set of experiments will examine effects of talker context. As introduced in Chapter 2, variation between talkers' voices can produce large differences in many acoustic cues (Jongman et al., 2000; McMurray & Jongman, 2011). In particular, one of the most salient sources of such differences is whether the talker is a man or woman. This results in a number of differences in acoustic cues, due to both articulatory factors and sociophonetic differences between men and women's speech. Moreover, these effects can be quite large. For example, gender accounted for approximately half of the variation in F1 frequency (one of the main cues to vowel quality) in a corpus of /ʌ/ and /ɛ/ vowels spoken by 10 talkers (Cole et al., 2010).

Variability due to talker gender is handled differently by the three types of cue encoding models. Raw-cue encoding approaches suggest that gender-specific acoustic differences are associated with phonological information at category levels. That is, to handle variation in vowel sounds across talker gender, listeners' vowel categories may be differentially activated depending on the gender of the talker. In contrast, the intrinsic and extrinsic encoding approaches both suggest that listeners' encode phonetic cues to vowels relative to the gender of the talker. However, these two approaches differ from each other in that intrinsic approaches use the relationship between cues within a single segment (e.g., the ratio between F1 and F2 Syrdal & Gopal, 1986), while extrinsic approaches use more global information (e.g., the size of a talker's vowel space Lobanov, 1971). Thus, speech perception in the presence of different talkers provides a platform for testing the predictions of these approaches to cue encoding, though previous work has not examined context effects at the level of cue encoding itself. The key distinctions between the models being examined here are illustrated in the differences between types of encoding in Figure 2.1.

To examine compensation for talker variability, I looked at differences in how listeners process VOT and F1, two cues that vary with talker gender. Both were examined as a function of whether a preceding carrier phrase was spoken by a man or woman. To remove the effects of gender information in the target word, stimuli with an ambiguous gender were used. Four experiments were conducted to determine how listeners compensate for differences in talker context and to examine vowel encoding using the N1/P3 approach used previously to examine encoding of a single cue, VOT, in the absence of context information (Toscano et al., 2010), with the N1 giving us a measure of cue encoding and the P3 a measure of categorization. In particular, an /ɛ/-/æ/ vowel contrast and a /b/-/p/ voicing contrast are used, since differences in talker gender affect cues to both of these distinctions.

Before presenting the experiments, a brief review of the relevant phonetic and perceptual data on talker identity effects will be presented, along with how the proposed compensation mechanisms have explained the data.

4.1.1   Phonetic data

As I introduced in Chapter 2, Peterson and Barney (1952) measured F1 and F2 values for English vowels from 76 talkers. They recorded the ten monopthong vowels in English in /hVd/ utterances (e.g., "hood", "had", "'hud", etc.) and found significant variation in formant frequencies for vowels spoken by men and women, reflecting differences in the size and shape of their articulators. Overall, women had higher F1 and F2 values than men. Moreover, there was considerable overlap across talkers for pairs of vowels that are close in F1xF2 space.

Hillenbrand et al. (1995) replicated the Peterson and Barney study and largely found the same results. One key difference was the properties of the vowels /æ/ and /ɛ/. Both studies found a great deal of overlap in steady-state F1 and F2 values between these two vowel sounds. However, Peterson and Barney found that /ɛ/ had lower F1 and F2 values than /æ/, while Hillenbrand et al. found lower F2 values for /æ/, and similar F1 values for

the two vowels (/æ/ had a slightly lower value for women, and /ɛ/ had a slightly lower value for men). In contrast, Di Benedetto (1989) found results consistent with Peterson and Barney's. Thus, these differences may have been due to factors specific to the talkers in the Hillenbrand et al. study, such as their dialect — they were mostly from southern Michigan and had the predominant dialect of that region. Indeed, a number of other studies have also found dialectical differences between vowel sounds (Clopper, Pisoni, & de Jong, 2005; Clopper & Pisoni, 2004; Hagiwara, 1997).

Regardless of the differences between the two datasets, both suggest that /æ/ and /ɛ/ are highly overlapping acoustically. They show that differences between talkers create situations in which the same vowel sound (as measured by F1 and F2) can be interpreted as two different vowels depending on whether it was spoken by a man or woman. For example, a given sound may be perceived as the vowel /æ/ (as in "had") when spoken by a man, and as /ɛ/ (as in "head") when spoken by a woman. Thus, it may be particularly useful for listeners to compensate for talker differences in these vowels.

The gender of the talker can also have effects on other acoustic features such as F1 intensity (Huber, Stathopoulos, Curione, Ash, & Johnson, 1999), formant frequency range (i.e., the size of the vowel space; Whiteside, 2001), and VOT. Swartz (1992) recorded men and women producing voiced and voiceless stop consonants. He measured VOT values for both groups and found that women have longer VOT values than men. As a result of this difference, the voicing category of intermediate VOT values can be ambiguous. Given gender information though, listeners may be able to compensate for this. Other studies have found similar results (Ryalls et al., 1997; Whiteside & Irving, 1998), though Morris, McCrea, and Herring (2008) found no effect when talkers produced syllables in isolation.

In addition, some gender effects are due to sociophonetic factors rather than articulatory differences. For example, spectral mean, one of the main acoustic cues for distinguishing place of articulation in English fricatives (e.g., /s/ vs. /ʃ/) varies between men and

women such that women have higher means than men (Strand & Johnson, 1996). Unlike the vowel effects discussed above, however, these effects appear to be driven by socio-cultural factors as a way to indicate the talker's gender. However, there is some evidence that listeners compensate for variation in this cue based on perceived vocal tract size (May, 1976), which could suggest an articulatory basis, though perceived vocal tract size would be a sociophonetic factor itself. Overall, the phonetic data suggest that a number of acoustic cues are affected by talker variability and that some kind of compensation for this during perception would be beneficial for listeners.

4.1.2 Perceptual data

Perceptual experiments suggest that listeners are sensitive to these phonetic relationships. Johnson et al. (1999) presented subjects with words spoken by men and women with either stereotypical or non-stereotypical voices (i.e., stereotypically-male for the men and stereotypically-female for the women). Auditory stimuli were also presented in the context of either a male or female face. They found that listeners' category boundaries along an F1 *hood-hud* vowel continuum shifted such that the boundary was lower in the context of a male talker. In addition, there was an effect of the visual face presented, again with a lower F1 boundary in the (visual) context of a male talker. This shift is consistent with the direction of compensation demanded by the phonetic data described above. Since men have lower formant frequencies on average, listeners treat an intermediate frequency as higher than it actually is, resulting in more /u/ responses and a lower F1 category boundary.

This result also demonstrates that compensation for talker gender need not be limited to auditory stimuli. Expectations driven by visual cues about the gender of the talker can serve a similar function. This rules out purely cue-level interactions (i.e., lateral interactions) as a source of the effects. Similarly, other talker-specific factors, such as dialect differences have an effect on vowel perception. Listeners are less able to discriminate vowel distinctions not present in their own dialect (e.g., *pin* vs. *pen*; Conrey, Potts, & Niedzielski,

2005) and categorize the same sounds differently depending on the dialect they were told the talker had (Niedzielski, 1999).

Talker gender also has perceptual effects on other speech contrasts. Strand and Johnson (1996) examined listeners' perception of fricative place of articulation (/s/ vs. /ʃ/) as a function of talker gender. Using a procedure similar to the one used by Johnson et al. (1999), they examined listeners' categorization along an /s/-/ʃ/ continuum varying in spectral mean as a function of whether the stimulus was spoken by a man or woman and whether a visually presented face was male or female. They found effects of both auditory and visual gender information, indicating that talker gender effects apply to fricative perception (a sociophonetic effect) as well as vowels.

There is also evidence that preceding sentential context can affect vowel judgments. Ladefoged and Broadbent (1957) presented listeners with synthesized sentences varying in the frequency range of their formants, simulating differences between talkers. Listeners were asked to make judgments about the last word in the sentence, which were of the form /bVt/ and differed based on their vowel. They found that listeners' responses to a given target word were influenced by the formant ranges in the carrier sentence. This suggests that compensation for talker variability occurs for temporally asynchronous stimuli in addition to stimuli in which talker and vowel information are heard (or seen) at the same time. This is relevant for the present study, since the effect of preceding context information will allow us to distinguish between different models of context compensation.

Together, these results allow us to rule out some of the proposed models of context compensation. Specifically, lateral, intrinsic encoding models cannot account for all of these results, since they do not predict effects of non-auditory information, sociophonetic differences, or vowel information in the preceding sentence. Thus, the main distinction the present set of experiments will focus on is the difference between extrinsic and raw-cue encoding models.

4.1.3   Mechanisms for handling talker variability

Several researchers have proposed the listeners' compensate for talker differences using an intrinsic encoding approach. Early work by Potter (1950) suggested that vowels are perceived based on the relative relationship between energy in different frequency bands, rather than the absolute values of formants, within a given segment. Presumably, this would allow listeners to handle differences between talkers' vowel productions. Syrdal (Syrdal, 1985; Syrdal & Gopal, 1986) implemented these ideas in a model of vowel categorization, and found that, across talkers, vowels clustered together better based on the distance in bark-space between formants (F2-F1) and F0 (F1-F0) than based on raw formant frequencies (F1 and F2). Fujisaki and Kawashima (1968) also found improved clustering using formant ratios, and listeners are sensitive to these differences (Christovich & Lublinskaya, 1979; Delattre, Liberman, Cooper, & Gerstman, 1952). However, as noted above, these models cannot account for previous results showing effects of sociophonetic differences, visual information, and the preceding sentence.

Ladefoged and Broadbent's (1957) results demonstrate that extrinsic information is also used to account for variability across talkers. They showed that preceding context can influence listeners' judgments, which argues against a purely intrinsic (e.g., formant ratio) account. Other researchers have suggested extrinsic normalization procedures based on compression or expansion of each talker's vowel space. Following up on Gerstman's (1968) early work using the range of formant frequencies for a given talker, Lobanov (1971) demonstrated that an algorithm that normalized a talker's vowel space based on the mean formant frequencies of their vowels produced much better clustering of vowel categories than raw formant frequencies. Since this relies on knowledge about a particular talker's vowel space (rather than the relationship between formants in an individual vowel), it fits into the category of extrinsic encoding via feedback models that suggest listeners recode formants on the basis of information about a particular talker. Similar effects are predicted

by C-CuRE (McMurray & Jongman, 2011; Cole et al., 2010), which uses expectations from various sources of context information to encode relative cues.

These accounts require highly-specific combinations of particular acoustic cues, namely steady-state formants. However, there is evidence that listeners can perceive vowels without steady-state formant cues at all. Strange, Jenkins, and colleagues (Strange, Jenkins, & Johnson, 1983; Jenkins & Strange, 1999; Jenkins, Strange, & Miranda, 1994) demonstrated this by presenting listeners with stimuli in which the steady-state vocalic portion had been removed. With these "silent-center" stimuli, listeners are still above chance at recognizing the correct vowel sound, even if the onset and offset are spoken by two different talkers (Jenkins et al., 1994). This result is most consistent with a raw-cue encoding account in which listeners are using cues other than steady-state formant frequencies (e.g., formant transitions, vowel durations) as cues to vowel quality. It is inconsistent with a strict intrinsic account that specifically requires listeners to recognize vowels on the basis of the relationships between formant frequencies or between formants and F0, though an intrinsic encoding process that incorporates changes over time may be able to account for these effects (Jenkins et al., 1994).

These studies suggest that both types of relative encoding, as well as raw-cue encoding approaches, can potentially play a role in vowel categorization, and some models have attempted to combine aspects of multiple approaches. For example, Nearey (1989) and Ainsworth (1975) demonstrate that both intrinsic and extrinsic factors can influence vowel categorization in a single experiment and, therefore, that both need to be incorporated into models of talker compensation. However, behavioral data alone make it difficult to assess whether this information is used at the level of cue encoding. Thus, the primary goal of this set of experiments is to examine cue encoding as a function of talker context to see whether listeners encode relative or raw cue-values.

4.1.4   Experiment overview and predictions

The studies described above demonstrate that there are a number of acoustic differences between talkers that listeners are sensitive to, particularly effects of talker gender. However, it is unclear which class of models ultimately underlies the way listeners deal with variability between talkers. Although purely intrinsic approaches are ruled out by the perceptual data, both extrinsic and raw-cue models, as well as those that incorporate aspects of intrinsic and extrinsic encoding, are possible.

To evaluate each of these classes of models, we need to examine the effect of preceding talker gender differences on cue encoding. This will allow us to contrast extrinsic encoding approaches with intrinsic and raw-cue encoding approaches and separate ERP responses to the context from those to the target stimulus. In turn, in order to do this we need to create stimuli in which the target word to be categorized has ambiguous gender information. In addition, we need to assess whether the ERP approach can allow us to detect encoding of other types of acoustic cues, such as formant frequencies. Thus, before examining the effect of talker context, these issues will be addressed in a series of experiments.

First, stimuli must have an ambiguous gender (otherwise listeners could simply use this information, rather than the sentence context, to make vowel judgments). The source-filter approach (i.e., linear predictive coding [LPC] resynthesis; Fant, 1960) can be used to generate stimuli that vary in pitch, formant frequency, and formant bandwidth to create male-female continua. Experiment 3 uses such stimuli to determine where listeners' category boundary is for male and female talkers. The ambiguously-gendered stimuli will then be used in subsequent experiments.

Second, we must establish that N1 amplitude can be used as an index of perceptual encoding for vowel distinctions. Experiment 1 established that differences in /æ/ and /ɛ/ have an effect on the N1, but it did not assesses within-category differences. Previous work has demonstrated, for voicing differences defined along VOT continua (Toscano et al., 2010),

N1 responses are continuous and do not reflect phonological differences. However, because the N1 does not appear to index perceptual encoding in general, but rather encoding for particular acoustic differences, we must establish that variation in N1 amplitude reflects continuous changes in the acoustic properties of the stimuli used here. In Experiment 4, the ambiguously-gendered stimuli from Experiment 3 will be varied along an F1 continuum from /ɛ/ to /æ/, and along a VOT continuum from /b/ to /p/. A target detection task will be used to examine effects of acoustic differences along the voicing and vowel continua on N1 and P3 amplitude.

Third, we must show that male and female carrier phrases will have an effect on listeners' categorization of the ambiguously-gendered stimuli. In Experiment 5, the ambiguously-gendered /ɛ/-/æ/ stimuli from Experiment 4 will be added to the end of a carrier phrase. By asking listeners to identify each stimulus we can examine the effect of talker context on vowel quality judgments. This allows us to test purely intrinsic approaches against extrinsic and raw-cue approaches, since the only gender differences occur outside the phonetic segments in the target words.

Finally, and most importantly, Experiment 6 addresses the primary question of interest: whether effects of talker context have an influence at the level of cue encoding or categorization. The stimuli from Experiment 5 will be used here, and as in Experiment 4, a target detection task will be used. Crucially, the target word will be preceded by a carrier phrase spoken by either a man or woman to set up expectations about gender.

This experiment directly tests the predictions of two of the encoding approaches. In particular, it allows us to compare the extrinsic approach (i.e., relative encoding from long-distance context information) with raw-cue encoding approaches by examining whether N1 responses to the onset of the target word vary as a function of the preceding carrier phrase.

Since women have longer VOTs than men, listeners would compensate for this by recoding a given sound as having a shorter VOT than it actually does. In addition, since the

results of Toscano et al. (2010) show that N1 amplitude decreases with increasing VOT, this would result in a larger N1 (i.e., consistent with a shorter VOT) in the context of a female talker than in the context of a male talker. Thus, if preceding context affects cue encoding, we would predict a smaller N1 in the context of a male talker than a female talker for stimuli varying in VOT.

The opposite pattern of results is predicted for the vowel stimuli. Since men have lower F1 frequencies than women, listeners would compensate by encoding a given F1 value as higher than it actually is when it is spoken in the context of a male talker. Given that the /æ/ stimuli in Experiment 1, which have higher F1 values than the /ɛ/ stimuli, produced larger N1s, we would expect larger N1s in the context of a male talker. Thus, relative encoding accounts predict that context should have different effects on N1 amplitude as a function of the particular acoustic cue (VOT or F1). This will help us to rule out alternative explanations based on overall differences in the N1 response to the different talkers.

Larger N1 amplitudes in the context of a female talker for the VOT stimuli and in the context of male talker for the F1 stimuli would demonstrate effects due to extrinsic encoding processes (i.e., the gender of the talker in the preceding sentence) and could not be accounted for by intrinsic encoding, since the relevant information needed for intrinsic compensation is constant across talker conditions. However, this result would also be consistent with accounts that allow for both extrinsic and intrinsic encoding. Raw-cue encoding approaches, in contrast, predict that cue encoding is not affected by context, but that categorization is. Thus, in this case, we would expect to see effects of talker gender in the P3 component and in listeners' overt responses, but not in the N1.

4.2 Experiment 3: Gender-neutral stimuli

Experiment 1 demonstrated that variation in naturally-produced /ɛ/-/æ/ sounds shows measurable effects on N1 amplitude. Given this, talker gender continua were created for *axe-ex* and *had-head* contrasts. Although listeners primarily use F0 to identify

talker gender (Gelfer & Mikos, 2005), both F0 and formant parameters (frequencies and bandwidths) were varied between male and female talkers to create natural-sounding endpoints. Vowel quality was also varied from /ɛ/ to /æ/ in three steps (the two endpoints and an ambiguous vowel) to see if listeners' gender judgments varied as a function of vowel. Listeners' categorized the stimuli as either male or female, and their category boundaries along each continuum were measured. In addition, ERPs were collected to see if differences in vowel quality for the manipulated stimuli had an effect on N1 amplitude (as the natural recordings used in Experiment 1 did).

4.2.1   Methods

4.2.1.1   *Participants*

Seventeen people participated in the experiment. Participant recruitment, consent, and compensation procedures were the same as in Experiment 1, and participants met the same language, hearing, and vision criteria as in previous experiments.

4.2.1.2   *Design*

Participants performed a 2AFC task in which they indicated whether an auditory stimulus was spoken by a male or female talker. Stimuli varied along two continua in nine gender steps and two vowel quality steps. Each stimulus was presented 10 times for a total of 540 trials. The experiment took approximately 60 minutes and was completed over the course of a single session.

4.2.1.3   *Stimuli*

The same tokens from the recordings made for Experiment 1 were used as the basis of the stimuli that were created for this experiment. For each selected word, a series of measurements were made using Praat. First, the onset and offset of steady-state energy in the vowel was measured. Second, pitch (F0) during the vowel was measured using the auto-
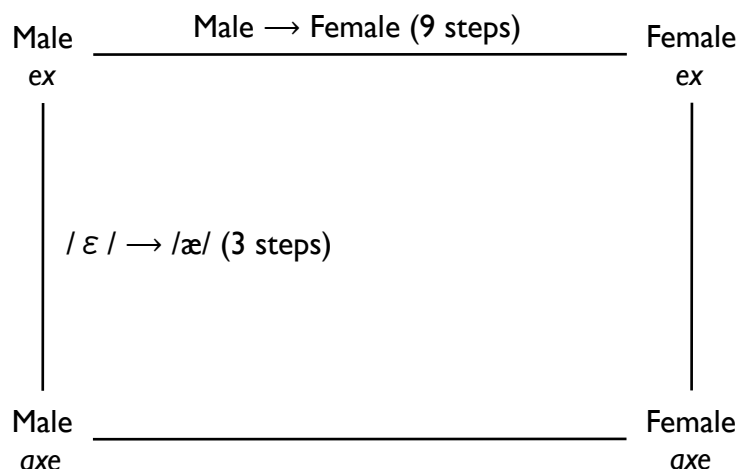
Figure 4.1: Schematic of Experiment 3 stimuli. For each word pair, stimuli varied from male to female in nine steps and from /ɛ/ to /æ/ in three steps.

mated pitch analysis tool in Praat with a time step of 0.01 s and a pitch floor of 75 Hz. Third, formants were measured using LPC with a time step of 10 ms, maximum formant frequency of 22.05 kHz, window length of 80 ms, and pre-emphasis from 214 Hz. The number of LPC poles used for each talker and word pair is given in Table 4.1. Mean formant frequencies across the vowel were used to measure the first four formants (F1-F4). Bandwidths for these four formants (B1-B4) were measured at the midpoint of the vowel.[1]

After these measurements were made, gender and vowel (/ɛ/-/æ/) continua were created using the *ex* and *head* tokens as starting points. Gender continua were made by gradually changing the parameters of the male token to match those of the female token. Nine parameters (F0, F1, F2, F3, F4, B1, B2, B3, and B4) were varied in nine steps along the gender continua and in three steps along the vowel continua. Figure 4.1 contains a schematic of the 27 stimuli for each word pair. Table 4.1 contains the starting and ending values for each parameter for the different continua.

---

[1]For the *ex* recording of the female talker, a different time point was chosen to measure the bandwidth of F3 due to a decrease in energy at that frequency range at the midpoint of the vowel.

Table 4.1: Endpoint values for Experiment 3 stimuli.

Gender	Word	LPC poles	F0	F1	F2	F3	F4	B1	B2	B3	B4
Male	*axe*	20	104	772	1831	2670	3245	195	269	436	869
	ex	20	104	671	1941	2709	3461	71	186	275	321
	had	17	104	733	1915	2807	3632	200	229	205	135
	head	17	109	541	2001	2837	3706	104	168	163	140
Female	*axe*	15	202	1056	1784	2853	4176	228	489	735	565
	ex	15	223	989	1972	2872	4119	183	347	438	439
	had	15	200	1065	1770	2983	4402	177	254	703	1065
	head	15	222	878	2046	3173	4362	66	343	782	762

Note: Formant frequency (F1-F4) and bandwidths (B1-B4) are in Hz.

The vocalic portion of each token was extracted, and F0 and formant values were varied using a two-stage procedure (thus, the vocalic portion was resynthesized twice). First, F0 was manipulated by extracting the pitch track (using the same parameters used for measurement), shifting it by the appropriate frequency for that step along the continuum (e.g., up 10 Hz), and replacing the pitch track of the sound. The sound was resynthesized using the pitch-synchronous overlap-add method (Moulines & Charpentier, 1990).

Next, LPC coefficients for the vowel were computed using 17 poles and the same parameters used for formant measurement. The sound source was then extracted from the original sound using the LPC coefficients. Formant frequencies and bandwidths were varied along continua by modifying the LPC coefficients, and the sound source was filtered through the modified formants to synthesize a new sound. The amplitude of the resynthesized sound was scaled to 0.99, resulting in similar intensities for each formant step.

After this, the pre- and post-vowel portion of each word from the male and female

tokens were sample averaged so that they had an ambiguous gender. That is, for *head*, the /h/ and /d/ from each token were averaged, and for *ex*, the /ks/ from each token was averaged. Prior to averaging, the sounds were equated for length by cutting the longer one at the zero-crossing closest to the duration of the shorter token. They were also normalized for intensity by setting the mean intensity of the lower-intensity sound to the mean intensity of the higher-intensity sound.

Finally, the sample averaged pre- and post-vowel parts of the word (the consonants) were then spliced onto the modified vowel. In order to remove high-frequency artifacts produced by the LPC procedure, the final stimuli were low-pass filtered to 11.125 kHz using a symmetric Hann filter with a smoothing width of 100 Hz.

4.2.1.4 *Procedure, EEG recording, and data processing*

The task was the same as in Experiment 1, except that participants categorized stimuli as male and female (indicated by the letters "M" and "F" on the screen). EEG recording and data processing procedures were the same as in Experiment 1.

4.2.2 Results

4.2.2.1 *Behavioral results*

Listeners' responses for both continua indicate that they reliably identified the stimuli as male or female at each endpoint (mean accuracy: 97.5%). Figure 4.2 shows participants' categorization responses as a function of continuum. The step closest to the point at which participants were equally likely to identify the stimuli as being spoken by a man or woman was step 4 for the *axe-ex* continuum and step 5 for the *had-head* continuum.

Behavioral responses were analyzed with a logit mixed-effects model with gender step, vowel step, and stimulus continuum entered as fixed effects (r_{max}=-0.466).[2] There was

---

[2]In the mixed-effects models reported here, the maximum correlation between the main effects in the model (r_{max}) is given as a measure of whether the results can be uniquely attributed to the relevant factors.
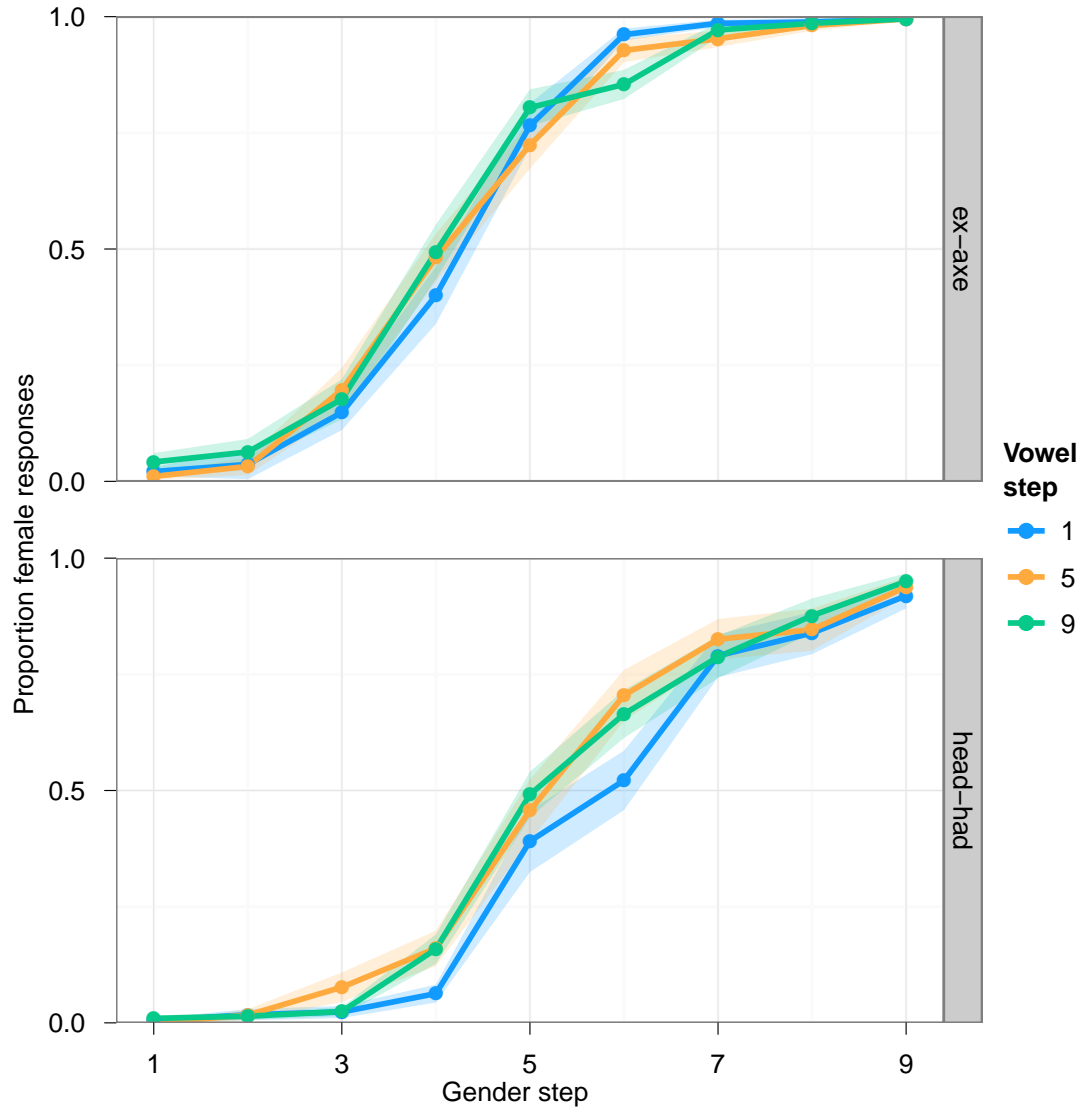
Figure 4.2: Experiment 3 results — behavioral responses. Participants' categorization responses as a function of gender step (male to female) for each vowel step (/ɛ/ to /æ/). The top panel shows listeners' responses for the *ex-axe* stimuli, and the bottom panel shows responses for the *head-had* stimuli.

a main effect of gender step (b=1.23, z=51.87, p<0.001) with more "female" responses at the continuum endpoint consistent with a woman's voice. There was also a main effect of continuum (b=-1.83, z=-25.18, p<0.001), which indicates that the category boundary differed between the two continua such that the boundary for the *ex-axe* continuum was closer to the *ex* endpoint than for the *head-had* continuum. The effect of vowel step was marginal (b=0.020, z=1.84, p=0.066).

There was also a gender step x continuum interaction (b=-0.34, z=-7.63, p<0.001), indicating that the *ex-axe* continuum had a steeper categorization function than the *head-had* continuum. There was also a gender step x vowel step interaction (b=-0.021, z=-3.04, p=0.002) and a gender step x vowel step x continuum interaction (b=0.040, z=2.95, p=0.003), indicating that listeners' gender categorization functions varied for the different vowel steps and that this effect differed for the two continua. This is not unreasonable, since the same set of acoustic cues that determines gender also determines vowel quality, and the stimuli were constructed to match the recorded tokens closely to make them as natural as possible. Thus, there could be particular combinations of gender and vowel step that produce similar response to other gender and vowel steps.

Finally, the vowel step x continuum interaction was significant (b=0.071, z=3.25, p=0.001). Separate models examining the effect of vowel step for the two continua were then run (*ex-axe*: r_{max}=-0.099; *head-had*: r_{max}=0.104). These revealed that that there was a significant effect of vowel step for the *head-had* continuum (b=0.060, z=4.22, p<0.001), but not for the *ex-axe* continuum (b=-0.017, z=-0.99, p=0.321). Listeners made more "female" responses for the *had* endpoint than for the *head* endpoint. Again, since the same acoustic cues signal differences in both vowel quality and gender, listeners may have used the overall combination of cue-values to make gender judgments.

In order to more precisely identify the male-female boundary for subsequent experiments, listeners' responses were also fit to four-parameter logistic curves, with the midpoint

Table 4.2: Category boundaries for Experiment 3.

Continuum	Vowel step	Category boundary
ex-axe	1	4.3
	5	4.4
	9	4.3
head-had	1	5.4
	5	5.8
	9	5.4

Note: Values are steps along each continuum.

of the function providing an estimate of the category boundary. The proportion of *female* responses as a function of gender step were fit for each subject, vowel step, and continuum. Overall, the mean boundary (averaged across the three vowel steps) was at step 4.3 for the *ex/axe* continuum and at step 5.5 for the *head/had* continuum, corresponding closely with the nearest steps where subjects' responses crossed the 50%-point for male/female responses. The mean boundaries for each continuum and vowel step are shown in Table 4.2.

4.2.2.2 ERP results

As in Experiment 1, ERP results for the two stimulus continua (*ex/axe* vs. *head/had*) were analyzed separately. For the *ex-axe* continuum, N1 amplitude was larger for the male endpoint than for the female endpoint along the gender continuum, and amplitude generally varied consistently with gender step across the continuum (with the exception of step 4). In addition, N1 amplitude was larger for the /æ/ endpoint than for the /ɛ/ endpoint, which fits with the results of Experiment 1. Figures 4.3 and 4.4 show grandaverage ERP waveforms for each word pair as a function of vowel step and gender step, respectively.

Mean N1 amplitude was measured for the average of the three frontal channels
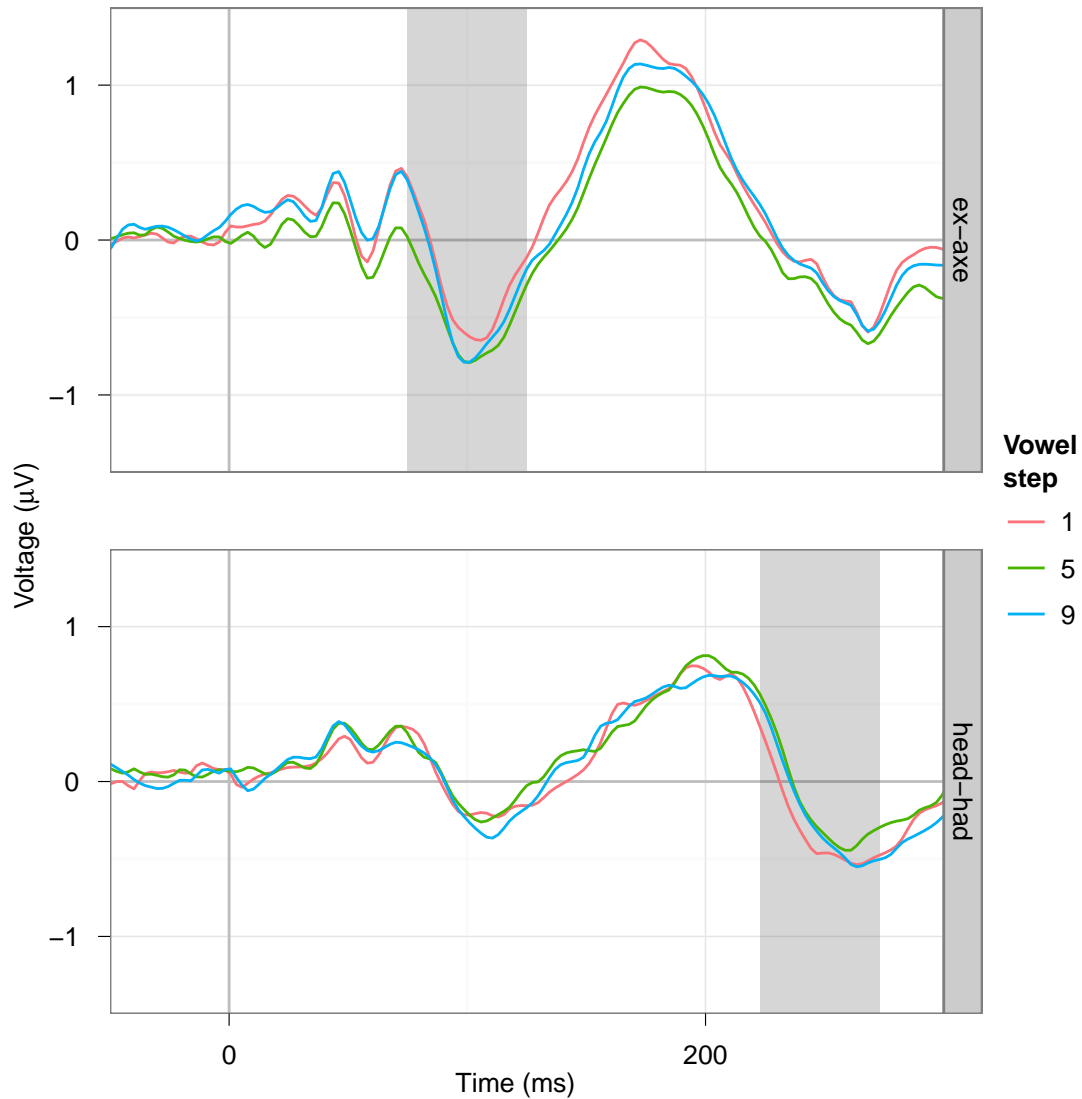
Figure 4.3: Experiment 3 results — ERP waveforms for vowel differences. Grandaverage waveforms for the average of the three frontal channels time-locked to word onset as a function of step along the vowel continuum for the *ex-axe* and *head-had* continua. Step 1 corresponds to the /ɛ/ endpoint and step 9 corresponds to the /æ/ endpoint. Shared areas represent the time range used to compute mean N1 amplitude (later for the *head-had* stimuli because of the word-intitial /h/).
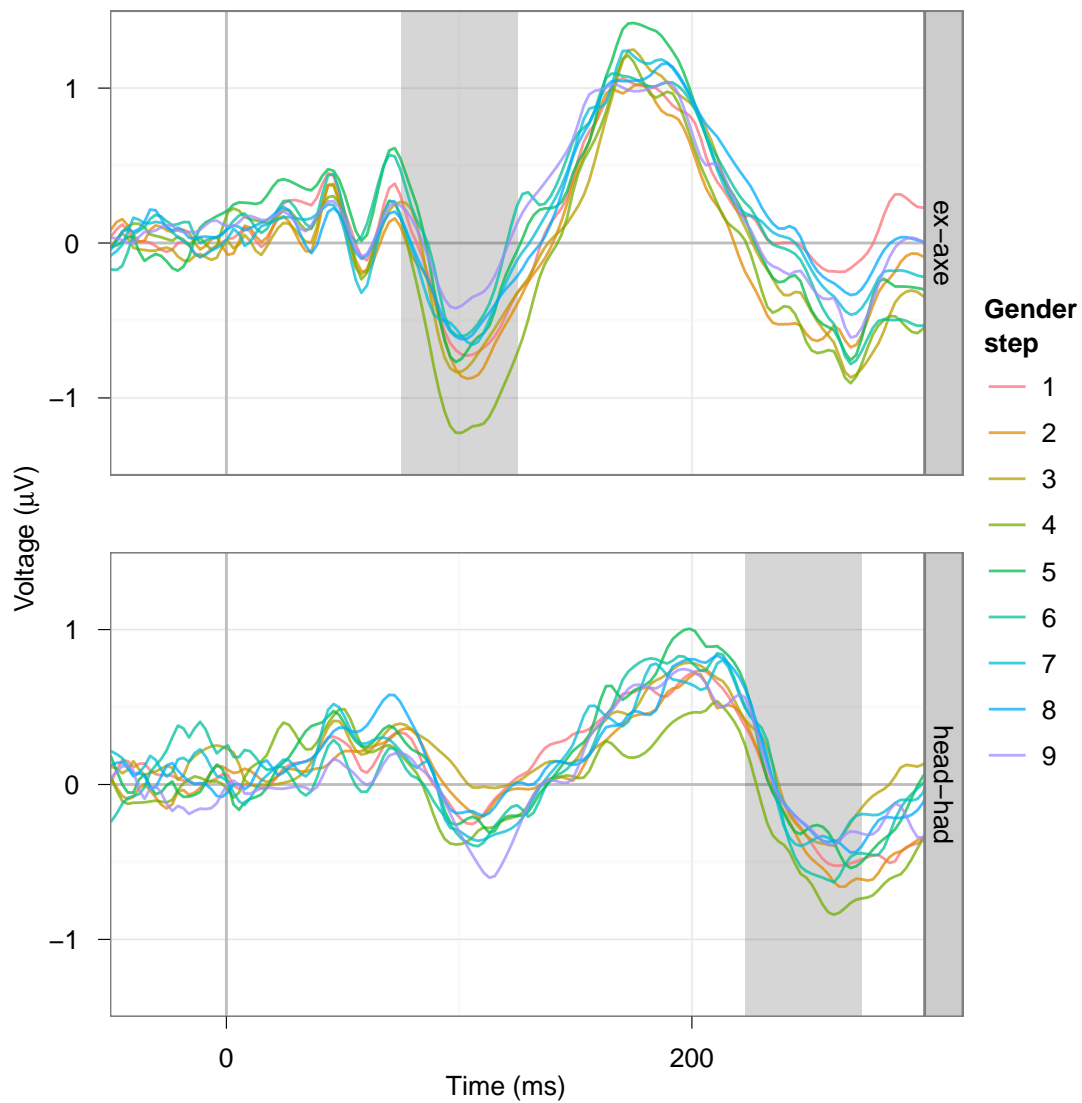
Figure 4.4: Experiment 3 results — ERP waveforms for gender differences. Grandaverage waveforms as a function of talker gender step for the two continua. Step 1 corresponds to the male endpoint, and step 9 corresponds to the female endpoint. Shared areas indicate time ranges used to compute mean N1 amplitude.

across the same time ranges after the onset of periodic energy used in Experiment 1 (75-125 ms for the *ex-axe* stimuli; 223-273 ms for the *head-had* stimuli). N1 amplitude as a function of gender and vowel step is shown in Figure 4.5. The mean amplitudes mirror the pattern of results in the waveforms. The figure also illustrates that the unusually large N1 amplitude for step 4 along the gender continuum was driven primarily by the ambiguous (i.e., step 5) /ɛ/-/æ/ stimuli.

N1 amplitude was analyzed using a linear mixed-effects model with gender and vowel step entered as fixed effects (r_{max}=0) in order to see if either showed a linear relationship with the N1.[3] There was a main effect of gender step (b=0.039, p_{MCMC}=0.009), indicating that N1 amplitude decreased linearly with increasing gender step. Although there was not a main effect of vowel step(b=-0.01, p_{MCMC}=0.307), this may have been due to variability in the ambiguous vowel stimuli and overlap from the P1 component.

There was a marginal gender step x vowel step interaction (b=0.009, p_{MCMC}=0.054), suggesting that the effect of gender differed as a function vowel step, though no clear pattern emerges from the data (see Figure 4.5). Given that variation in the same set of acoustic cues signals both gender and vowel differences, it would not be surprising if particular combinations of gender and vowel step were acoustically similar to other conditions, producing similar effects on N1 amplitude. However, it is difficult to determine precisely which cues are driving the effect in this case, since a large number were varied in an effort to create stimuli that sounded as natural as possible.

As in Experiment 1, there was little variation in N1 amplitude along the *head-had* continuum. Again, this may be due to overlap from ERP components to the initial /h/ in these stimuli. A linear mixed-effects model examining N1 amplitude (r_{max}=0) did not find an effect of gender (b=0.02, p_{MCMC}=0.283) or vowel step (b=0.01, p_{MCMC}=0.378), nor an

---

[3]In this and subsequent linear mixed-effects analyses, the Markov chain Monte Carlo (MCMC) procedure was used to estimate p-values (p_{MCMC}) for each effect, with 10,000 simulations being run in each case.

interaction (b=0.003, p_{MCMC}=0.483). Again, this is likely due to overlap from the preceding /h/. Partially because of this, subsequent experiments used only the *ex-axe* continuum.

4.2.3   Discussion

The results of this experiment suggest that the LPC resynthesized stimuli can be correctly perceived as being spoken by a man or woman, and more importantly, that by manipulating pitch, formant frequencies, and formant bandwidths, we can create gender-neutral stimuli. The parameters of the stimuli at the category boundaries in this experiment will be used for Experiments 4-6.

The results also suggest that simultaneously manipulating multiple cues to vowel quality can make it difficult to observe changes in the N1 as a function of acoustic differences. In the next experiment, only F1 will be manipulated along the /ɛ/-/æ/ continuum. This is one of the two cues (along with F2) that distinguished these two vowels in the Peterson and Barney (1952) dataset. In addition, because neither this experiment nor Experiment 1 found any large differences in N1 amplitude for the *head-had* continuum, only the *ex-axe* continuum will be used in subsequent experiments.

4.3   Experiment 4: ERP responses to vowel continua

The results Toscano et al. (2010) suggested that N1 amplitude can be used as an index of perceptual encoding, since it varies linearly with changes in VOT and is not influenced by participant's phonological categories. However, it is not clear whether this also applies to other phonetic cues, such as the formant frequencies manipulated in this set of experiments. To create stimuli that vary in only a single acoustic cue, F1 was manipulated for the /ɛ/-/æ/ stimuli used here. Thus, it is important to confirm that the same effect can be observed for differences in F1.

This experiment is designed to establish that N1 amplitude varies linearly and independently of listeners' phonological categories as a function of changes in F1 and to replicate
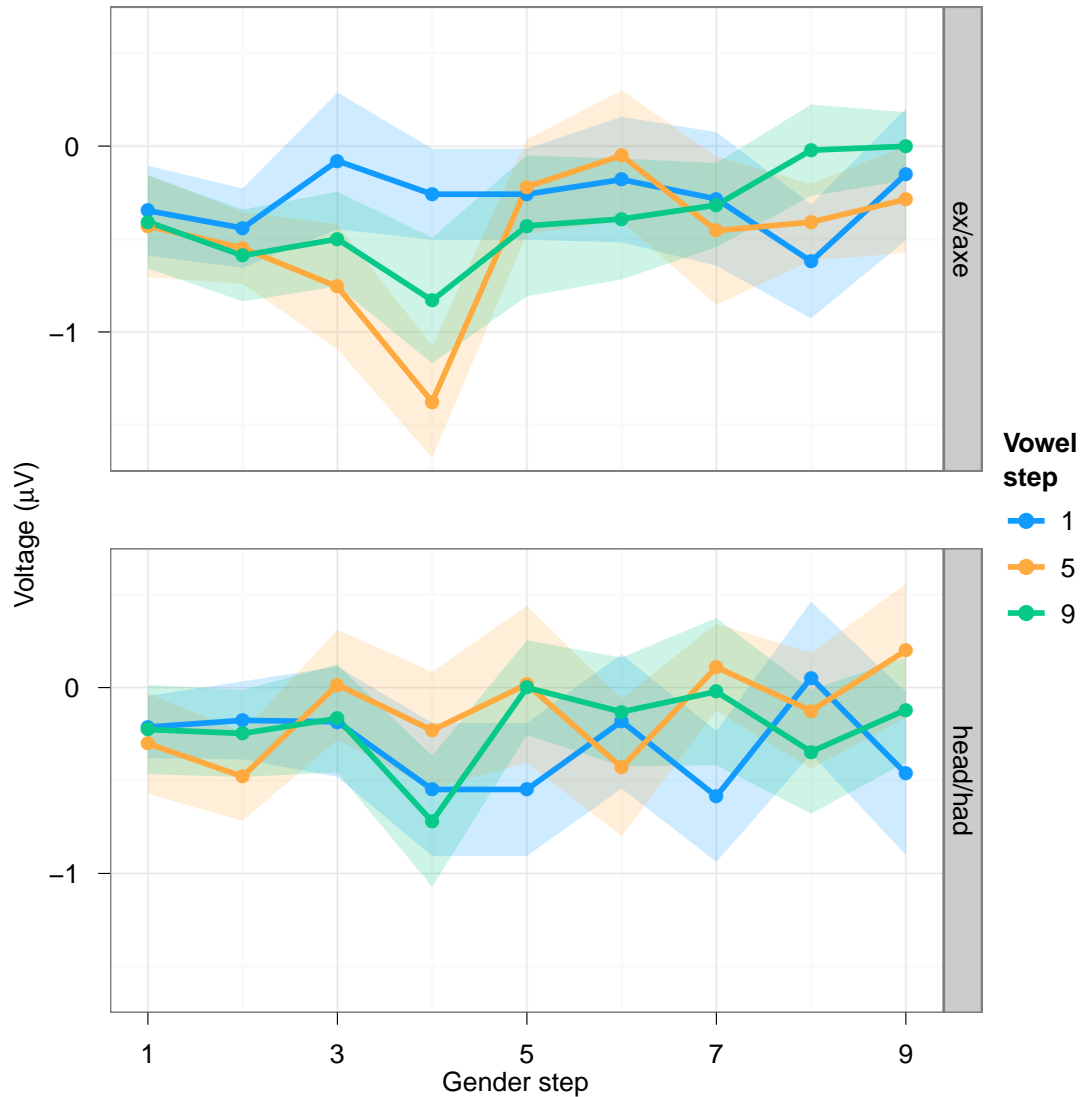
Figure 4.5: Experiment 3 results — N1 amplitude. Mean amplitudes for the average of the three frontal channels as a function of gender and vowel step. N1 amplitude was calculated as the mean voltage 75-125 ms after the onset of the vocalic portion. For the *ex-axe* continuum, this occurred from 75-125 ms post-stimulus, and for the *head-head* continuum, from 223-273 ms post-stimulus (accounting for the word-initial /h/).

the results of Toscano et al. (2010) for VOT. Listeners' heard stimuli varying in F1 along an *ex-axe* continuum (increasing F1 from the /ɛ/ to the /æ/ endpoint), as well as stimuli varying in VOT along a *bee-pea* continuum (increasing VOT from the /b/ to /p/). These two continua will also be used in the experiment examining the effect of preceding gender context, since they make different predictions about the direction of the effect: relative encoding approaches predict a larger N1 in the context of a male talker for the *ex-axe* stimuli, and a larger N1 in the context of a female talker for the *bee-pea* stimuli.

The ERP data were analyzed using approaches similar to those described in Toscano et al. (2010). Two criteria were used to assess whether N1 differences reflect cue encoding. First, to examine the overall effect of F1 and VOT on N1 amplitude, the data from each stimulus continuum were analyzed using a linear mixed-effects model with step and target as fixed effects. A significant effect of step in this analysis would demonstrate a linear relationship between the cue and N1 amplitude, providing the first piece of evidence that it reflects continuous cue encoding. Second, this analysis was repeated with the data grouped by subjects' responses to rule out the possibility that linear effects in the first analysis were due to averaging across categorical differences in N1 amplitude. In this analysis, a main effect of subjects' response with no effect of continuum step or interaction would suggest that differences in N1 amplitude reflect a categorical distinction between the stimuli rather than a continuous effect. Further, a significant effect of step would provide additional evidence that the N1 reflects continuous cue encoding.

Next, to examine whether the P3 reflects variation within a phonological category, two analyses were conducted. First, differences in P3 amplitude were analyzed as a function of distance from the target endpoint. Stimuli at the endpoint should produce the largest P3 and within-category differences in P3 amplitude would suggest that listeners' are sensitive to acoustic differences within a phonological category at late stages of processing. Second, to establish that these differences reflect effects based on individual listeners' categories, the

data were analyzed as a function of distance from each listener's category boundary for the target-response trials along that cue dimension.

Together these analyses allow us to determine whether the N1 reflects encoding of continuous acoustic differences and whether the P3 reflects category-level differences for the two continua used here.

4.3.1 Methods

4.3.1.1 *Participants*

Twenty-seven people participated in the experiment. Participant recruitment, consent, and compensation procedures were the same as the other ERP experiments, and participants met the same language, hearing, and vision criteria as in previous experiments. One participant was excluded from the analyses for the *bea-pea* stimuli, and four were excluded from the *ex-axe* analyses due to poor performance (<50% correct) on the endpoints.

4.3.1.2 *Design*

The design of this experiment is similar to the one used by Toscano et al. (2010) to measure the N1 and P3 components. Participants performed a target detection task in which they identified whether stimuli matched a target word. Stimuli consisted of nine-step continua varying in VOT from *bee* to *pea* and in F1 from *ex* to *axe*. Each word served as the target in separate parts of the experiment. Each stimulus was repeated 15 times in random order, for a total of 1140 trials.

Because the target detection task leads to different category boundaries as a function of the target category (Toscano et al., 2010), an additional 2AFC task was used to obtain participants category boundaries for the target words in each stimulus continuum. This was run after the EEG recording session. Stimuli were blocked by continuum, and the order of the continua and mapping between the endpoints and left and right buttons was randomized between participants. Participants heard five repetitions of each stimulus. The

two parts of the experiment took approximately two hours and were run in a single session.

4.3.1.3 Stimuli

The *ex-axe* stimuli were based on the same set of recordings as in Experiment 3, and the same procedures were used to create the vowel continua, except that nine steps were used instead of three and only the F1 parameter was varied to create the vowel continuum. First, the acoustic parameters corresponding to the gender of the talker were set to the ambiguous values obtained from the results of Experiment 3. Pilot categorization data revealed that the F1 endpoint values based on the measurements of the original recordings did not produce reliably correct responses at the continuum endpoints. Thus, the range of values used was increased slightly so that the /ɛ/ and /æ/ endpoints would be more clearly identified; F1 values varied over a 163 Hz range across the continuum. Table 4.3 shows the F1 values for each step along the /ɛ/-/æ/ continuum.

For the *bee-pea* stimuli, a VOT continuum was generated using the same cross-splicing procedure used for Experiment 2. Pilot data indicated that listeners reliably identified the /b/ and /p/ endpoint steps correctly. VOT values are shown in Table 4.4.

4.3.1.4 Procedure

The experimental setup was the same as in Experiment 1. The fixation point and button labels on the screen were replaced by two lines of text indicating the target word on the top line and the corresponding button on the bottom line. The target button (left or

Table 4.3: F1 values for Experiment 4 /ɛ/-/æ/ continuum steps.

Continuum	1	2	3	4	5	6	7	8	9
ex-axe	727	749	773	797	817	833	851	868	890

Note: Values are in Hz.

right) was constant for each participant and alternated between participants.

Each individual stimulus was equally likely. As a result, targets occurred on approximately 25% of the trials in each block (depending on where the participant's category boundary was for those stimuli), which should produce a P3 response. Target block order was varied between participants with the restriction that words from the same stimulus continuum (*bee/pea* or *ex/axe*) could not serve as the target in adjacent blocks.

4.3.1.5   *EEG recording and data processing*

EEG and data processing procedures were the same as in the previous experiments.

4.3.2   Results

4.3.2.1   *Behavioral responses*

Most participants correctly identified the stimulus endpoints in each target block. Overall mean accuracy at the continua endpoints was 93.4% including all participants, and 95.8% including only those with greater than 50% correct responses on the endpoints. Figure 4.6 shows the these listeners' responses for each stimulus continuum as a function of continuum step and target. As expected, listeners' categorization functions were shifted as a function of which endpoint was the target (i.e., they showed a bias toward making more target responses).

Logit mixed-effects models were used to analyze listeners responses for each stimulus continuum for the two target blocks relevant for that continuum (e.g., listeners responses

Table 4.4: VOT values for Experiment 4 /b/-/p/ continuum steps.

Continuum	1	2	3	4	5	6	7	8	9
bee-pea	0	5	12	19	20	24	30	32	39

Note: Values are in ms.
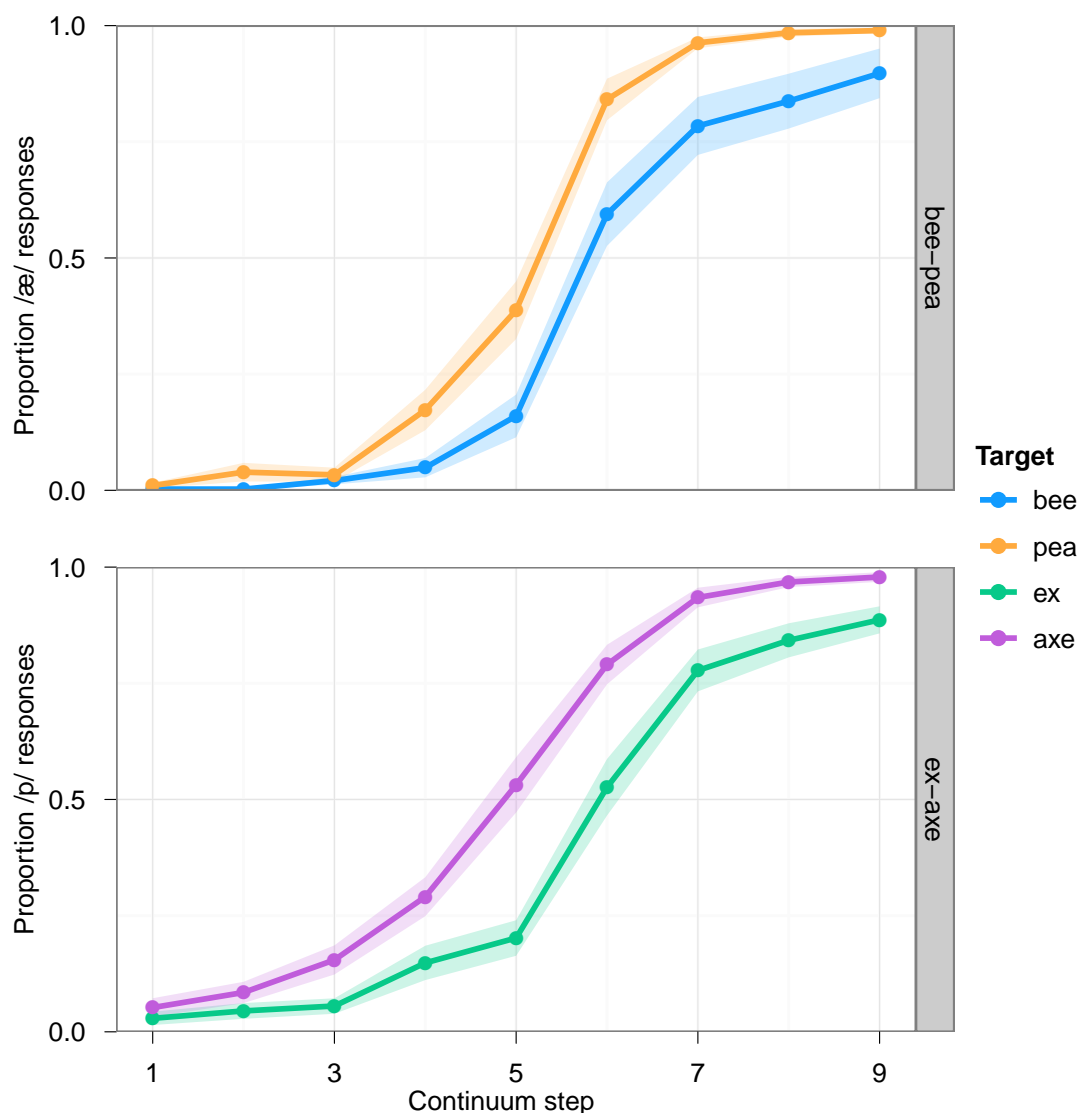
Figure 4.6: Experiment 4 results — behavioral responses in target detection task. (Top panel) Proportion of *pea* responses for *bee-pea* stimuli in the *bee* and *pea* target blocks. (Bottom panel) Proportion of *axe* responses for *ex-axe* stimuli in the *axe* and *ex* target blocks. For both continua, listeners' category boundary shifted such that they made more "target" responses (i.e., more *pea* responses when *pea* was the target than when *bee* was the target).

to *bee-pea* stimuli when either *bee* or *pea* served as the target). Step and target were entered as fixed effects. For the analysis, target/non-target responses were converted to responses relative to the step-9 endpoint for each continuum. For example, for a /b/-/p/ stimulus with *pea* as the target, a "target" response was coded as a /p/ and a "nontarget" response was coded as a /b/; when *bee* was the target, a "target" response was coded as /b/ and a "non-target" response was coded as /p/.

For the *bee-pea* stimuli, a model with VOT step and target word (*bee* vs. *pea*) was run (r_{max}=0.36). There was a main effect of step (b=1.09, z=49.00, p<0.001), indicating that listeners responses varied as a function of VOT. There was also a main effect of target (b=0.47, z=14.98, p<0.001), which indicates that listeners produced more /p/ responses when *pea* was the target than when *bee* was the target. The interaction was also significant (b=-0.16, z=-9.06, p<0.001), indicating that the slope of the categorization function was significantly greater for the *pea* target block than for the *bee* target block.

For the *ex-axe* stimuli, a model with F1 step and target word (*ex* vs. *axe*) was run (r_{max}=0.13). There was also a main effect of step (b=1.14, z=51.28, p<0.001), target (b=0.21, z=4.60, p<0.001), and step x target interaction (b=-0.16, z=-6.03, p<0.001), the same pattern of results seen with the *bee-pea* stimuli.

Listeners' responses during the 2AFC task run after the ERP session also showed standard categorization functions. As in Experiment 2, these data were fit to four-parameter logistic functions to determine participants' category boundaries (these will be used in the P3 analyses below). The mean boundary for both (9-step) continua was at step 5.4.

4.3.2.2 N1 amplitude

As in previous experiments, data from each stimulus continuum was analyzed separately, since the two continua varied along two different acoustic cue dimensions and ERPs to different steps along each continuum are not directly comparable. Figure 4.7 shows grandaverage ERP waveforms for the average of the three frontal channels as a function of

continuum step for each of the two stimulus continua.

For the *bee-pea* continuum, the latency of the N1 was later and appeared to vary more with VOT than in previous experiments. Because of this, the previous time range used to compute mean N1 amplitude (75-125 ms) might correspond to the N1 for some VOTs but not others. Thus, N1 amplitude was computed from the mean voltage 100 to 140 ms post-stimulus, a time range that included the N1 across the VOT continuum. For the *ex-axe* continuum, the same 75-125 ms post-stimulus time range used in previous experiments was used. Figure 4.8 shows mean N1 amplitude as a function of step and target type (*ex-axe* vs. *bee-pea*) for each of the stimulus continua.

For the *bee-pea* stimuli, short VOTs produced larger N1s than long VOTs, as in previous experiments. Mean N1 amplitude was analyzed using a linear mixed-effects model with VOT step and target type as fixed effects (r_{max}=0).[4] The model showed a main effect of VOT step (b=0.028, p_{MCMC}=0.014), which replicates the results of Toscano et al. (2010) showing that N1 amplitude decreases linearly with increasing VOT. There was not an effect of target type (b=-0.083, p_{MCMC}=0.157) nor an interaction (b=0.005, p_{MCMC}=0.827).

Next, N1 amplitude was analyzed as a function of listeners' responses. To look at this for the *bee-pea* stimuli in this experiment, trials on which listeners made a "target" response were analyzed as a function of whether the target was *bee* or *pea*. Because this leads to many more trials for one condition than the other (i.e., there are very few trials in which listeners responded "target" at the *bee* endpoint when *pea* was the target), each condition was weighted by the number of trials in it. Figure 4.9 shows mean N1 amplitude as a function of continuum step for the target-response trials for each stimulus continuum. The weight (i.e., the number of trials) in each condition is indicated by the relative sizes of the data points in the figure.

---

[4]Note that while we separated the data by stimulus continuum, the design of the experiment was such that when the subject *heard* stimuli from the F1 (*ex-axe*) continuum, they would still be monitoring for *bee* or *pea* on 50% of the trials
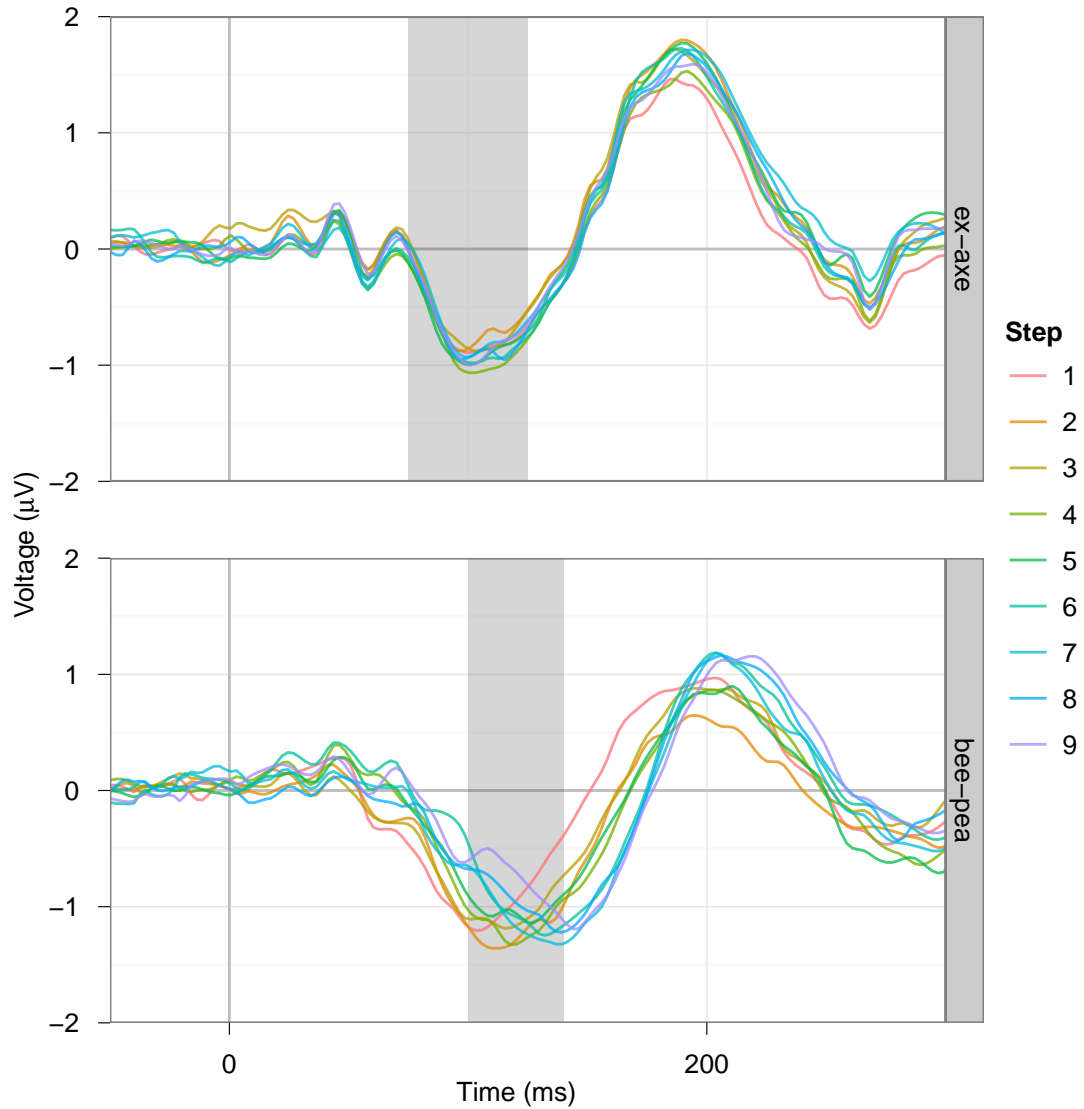
Figure 4.7: Experiment 4 results — ERP waveforms for frontal channels by continuum step. The top panel shows ERPs to the *ex-axe* stimuli, and the bottom panel shows ERPs to the *bee-pea* stimuli. For the *bee-pea* stimuli, N1 amplitude decreases with increasing VOT. Shaded areas indicate time ranges used to compute mean N1 amplitude.
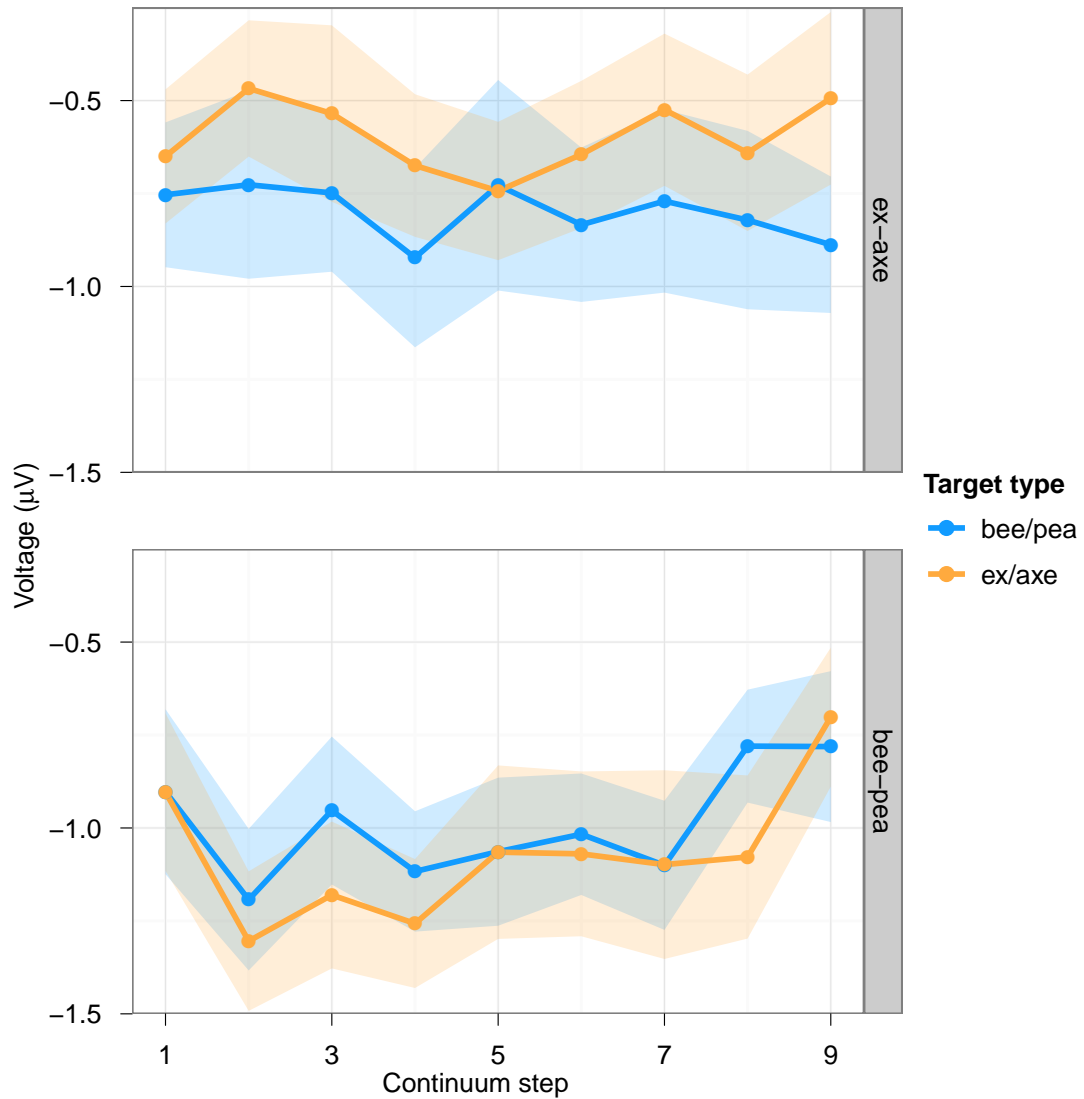
Figure 4.8: Experiment 4 results — N1 amplitude. Mean amplitude as a function of continuum step and target type (*bee/pea* vs. *ex/axe*) for each stimulus continuum (top panel: *ex-axe*; bottom panel: *bee-pea*).

A weighted linear mixed-effects model with step and target (*bee* vs. *pea*) as fixed effects was run (r_{max}=-0.737).[5] The model showed a marginal effect of VOT step (b=0.046, p_{MCMC}=0.083), and no effect of target (b=0.073, p_{MCMC}=0.748) or interaction (b=-0.047, p_{MCMC}=0.501). Thus, there is no indication that N1 amplitude varied as a function of which endpoint served as the target, though the overall linear trend across the VOT continuum was not as strong as in the originally analysis. This may have been partially due to the reduced number of trials in this dataset.

Mean N1 amplitude for the *ex-axe* stimuli were analyzed using the same approach. First, a linear mixed-effects model was used to examine N1 amplitude in the overall dataset. F1 step and target type (*ex/axe* vs. *bee/pea*) were entered as fixed effects (r_{max}=0). The effect of F1 step was not significant (b=-0.005, p_{MCMC}=0.626). However, given that the effect of F1 on the N1 is relatively small, it may be difficult to detect a linear trend for these stimuli.

The model did show a main effect of target type (b=0.20, p_{MCMC}<0.001), with the *bee/pea* targets producing larger N1s than the *ex/axe* targets. The interaction between F1 step and target type was not significant (b=0.02, p_{MCMC}=0.470). The effect of target type could be due to differences in attention to the stimuli in the different target conditions, though increased attention typically produces larger N1s (Hansen & Hillyard, 1980), and it is unclear how listeners' are attending to the stimuli in the target detection task used here. However, there are other conditions in which additional information that is relevant to the speech produces smaller N1s (e.g., audiovisual speech, van Wassenhove et al., 2005). The effect of target type here may reflect a similar process.

Although an overall effect of F1 step was not found, an effect may be apparent if we control for listeners' responses. This analysis was not significant for the VOT stimuli in this experiment, but Toscano et al. (2010) did find an effect of VOT when listeners' responses

---

[5]Because this model is highly unbalanced (e.g., since there are many more "target" responses to the shorter VOTs when *bee* is the target), we would expect that the correlations between fixed effects may be relatively large.
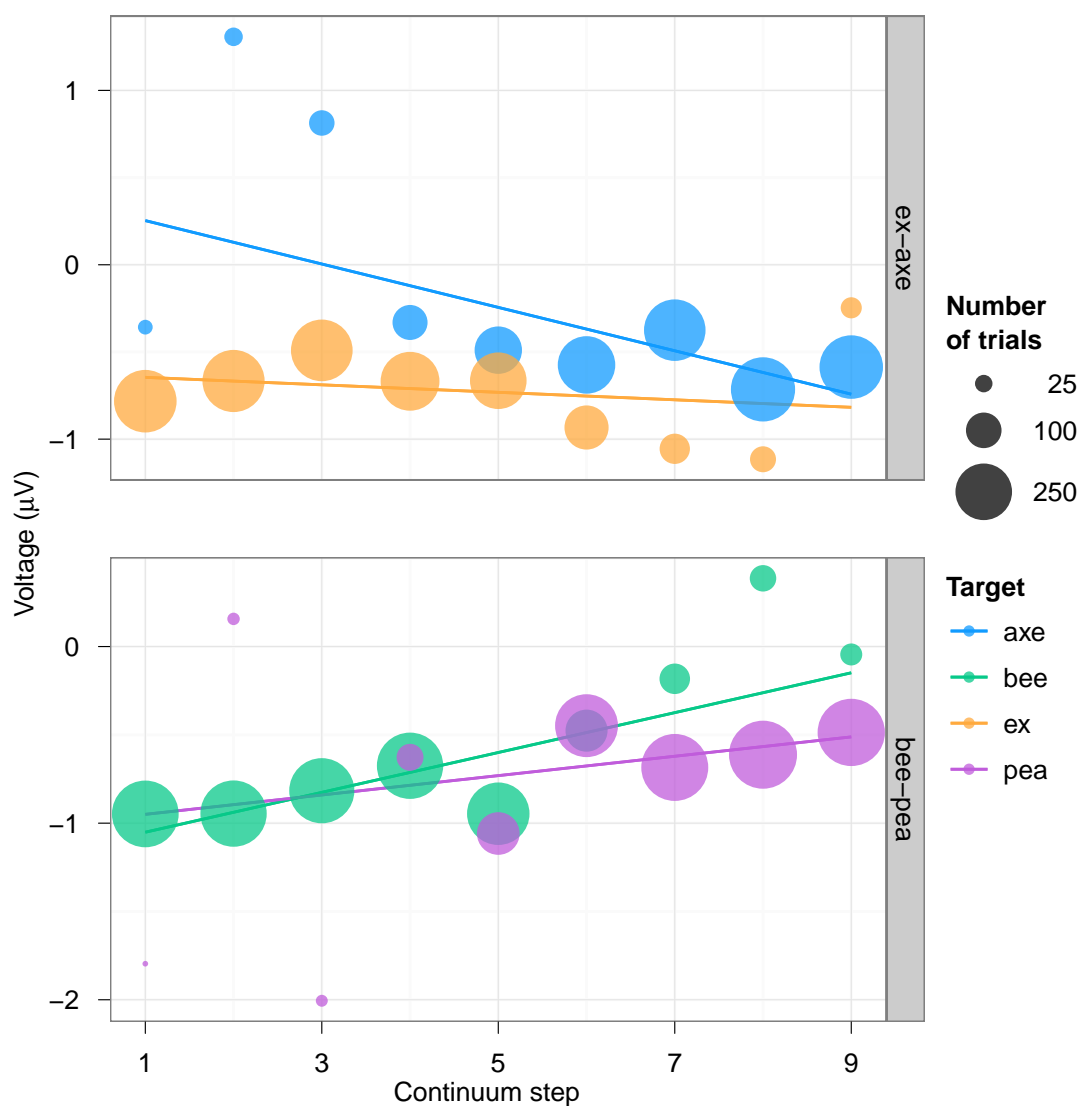
Figure 4.9: Experiment 4 results — N1 amplitude on target-response trials. Mean amplitude as a function continuum step, target word, and stimulus continuum for trials on which the participant made a "target" response. The top panel shows ERPs for the *ex-axe* stimuli, and the bottom panel shows ERPs for the *bee-pea* stimuli. Because listeners made few "target" responses to stimuli at the non-target endpoint (e.g., a 0 ms VOT [*bee*] when *pea* is the target), the size of each data point is weighted by the number of trials in that condition. Lines represent weighted linear models.

were taken into account by looking at only the "target"-response trials. This suggests that N1 amplitude is at least partially related to listeners' ultimate responses. In addition, given the variability in listeners' responses to these stimuli, an effect of F1 on N1 amplitude may be more apparent for trials on which listeners made the same response.

Thus, as with the *bee-pea* stimuli, trials on which listeners made a "target" response were analyzed as function of continuum step and target (*ex* vs. *axe*). A weighted linear mixed-effects model with F1 step and target (*ex* vs. *axe*) was run (r_{max}=-0.640). The model showed a main effect of F1 step (b=-0.059, p_{MCMC}=0.024) such that N1 amplitude increased with increasing F1 (i.e., closer to the /æ/ endpoint). This mirrors the results of Experiment 1 showing larger N1s for the /æ/ stimuli. There was also a main effect of target (b=0.38, p_{MCMC}=0.002) with larger N1s when *ex* was the target than when *axe* was the target. The interaction was also significant (b=-0.10, p_{MCMC}=0.047). A follow-up analysis examining each target condition separately found a significant effect for the *axe*-target trials (b=-0.11, p_{MCMC}=0.006), but not for the *ex*-target trials (b=-0.001, p_{MCMC}=0.969). This suggests that the N1 is linearly sensitive to changes in F1, though the effect was only apparent for the *axe*-target trials. Critically, this did not interact with the effect for the *ex*-target trials such that the direction of differences in N1 amplitude varied depending on listeners' responses (and which word they were monitoring for). Thus, there is no evidence to indicate an influence of phonological category information on the N1 response.

4.3.2.3 P3 amplitude

P3 amplitude was analyzed as a function of continuum step and target type for each stimulus continuum. Because we expect the P3 to be largest for the (infrequent) target stimuli, it should be largest at the relevant continuum endpoint, and, if it is sensitive to variation within a phonetic category, to decrease with distance from that endpoint. In addition, we do not expect a P3 when one of the words from the opposite stimulus continuum is the target.

Figures 4.10 shows grandaverage waveforms for the average of the three parietal

channels as a function of trial type (target vs. nontarget; e.g,. the *ex* and *axe* target blocks are "target" trials for the *ex-axe* stimuli and "nontarget" trials for the *bee-pea* stimuli). The figure illustrates that, as expected, P3 amplitude was larger for the target than the nontarget trials. Figure 4.11 shows ERP waveforms for the target trials as a function of distance from target endpoint. For both continua, P3 amplitude was largest at the relevant target endpoint (e.g., the step 9 [high F1] endpoint when *axe* is the target), and a much smaller P3 was elicited by the other endpoint (e.g., step 1). Figure 4.12 shows mean P3 amplitude (average voltage from 300 to 700 ms post-stimulus) for the target trials from each stimulus continuum as a function of target endpoint distance.

To examine whether the target trials produced larger P3s than the nontarget trials for the *bee-pea* stimuli, a linear mixed-effects model was run with trial type (target [*bee/pea*] vs. nontarget [*ex/axe*]) as a fixed effect. The model showed a significant effect (b=0.99, p_{MCMC}<0.001), indicating that P3 amplitude was greater for the *bee* and *pea* target blocks than for the *ex* and *axe* target blocks. Next, a model analyzing the target trials was run with distance from the target endpoint as a fixed effect.[6] There was also a main effect of target distance (b=-0.086, p_{MCMC}<0.001), indicating that P3 amplitude decreased with increasing distance from the target endpoint. This confirms the prediction that listeners are sensitive to within-category acoustic differences at post-perceptual stages.

The data were also analyzed as a function of each individual participant's category boundaries along the *bee-pea* continuum for trials in which they made a "target" response on the *bee* and *pea* target blocks. This allows us to confirm that P3 amplitude varies within each listener's phonetic categories for stimuli that were classified the same for a given target. First, target endpoint distance was recoded relative to each participant's category boundary (computed from the 2AFC task run after the ERP session; see section 4.3.2.1 above) such that negative values indicate steps in the non-target phonetic category, and positive values

---

[6]Two separate models (one examining the effect of trial type, and one examining the effect of target distance) were run because target distance isn't defined for the nontarget trials.
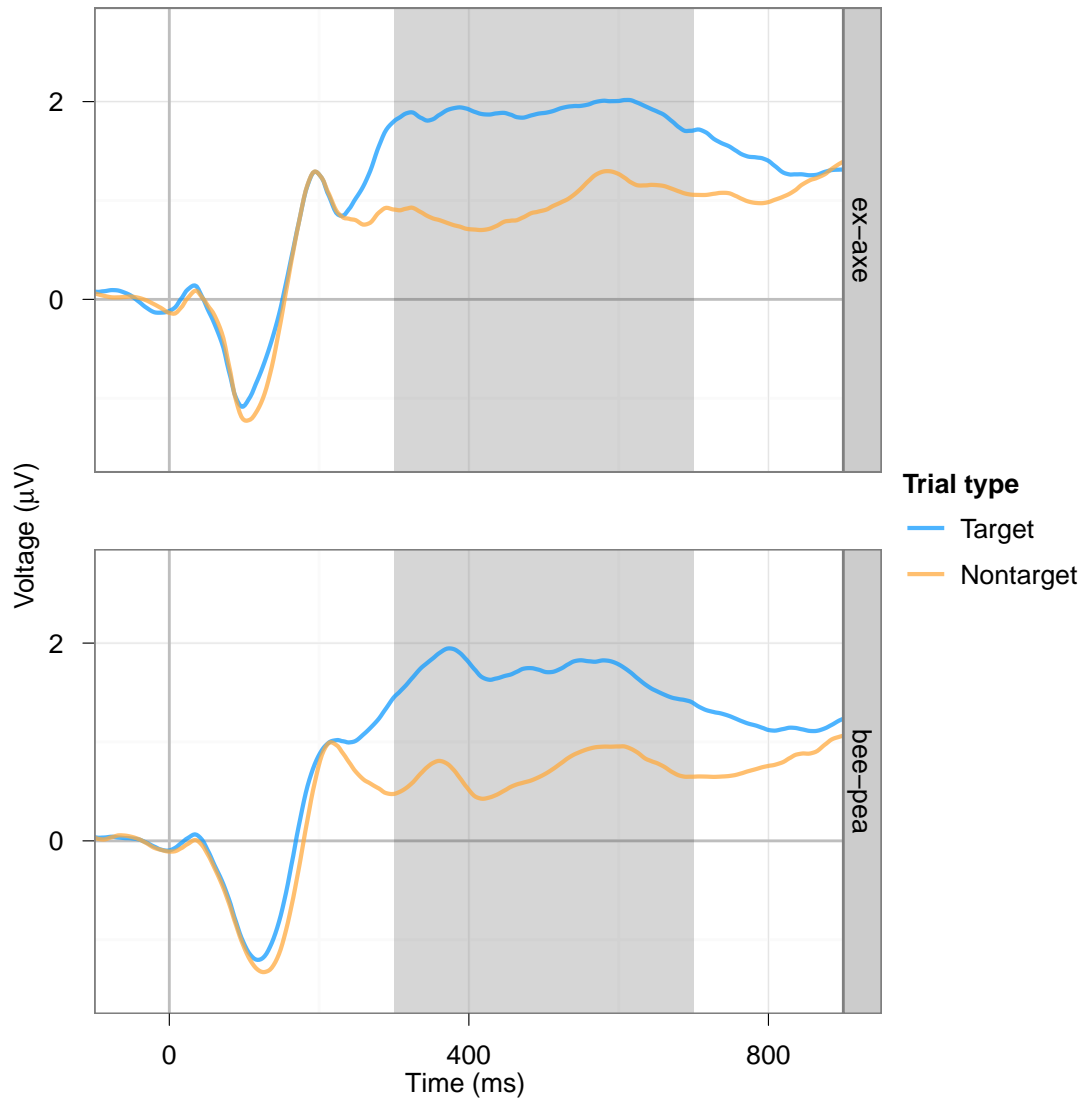
Figure 4.10: Experiment 4 results — ERP waveforms for parietal channels by trial type. For the *ex-axe* stimuli (top panel), the *ex* and *axe* target blocks are the "target" trials and the *bee* and *pea* target blocks are the "nontarget" trials. For the *bee-pea* stimuli (bottom panel), the *ex/axe* blocks are "nontarget" trials and the *bee/pea* blocks are "target" trials. P3 amplitude is greater for the target trials than the nontarget trials. Shaded areas indicate the time range used to compute mean P3 amplitude.
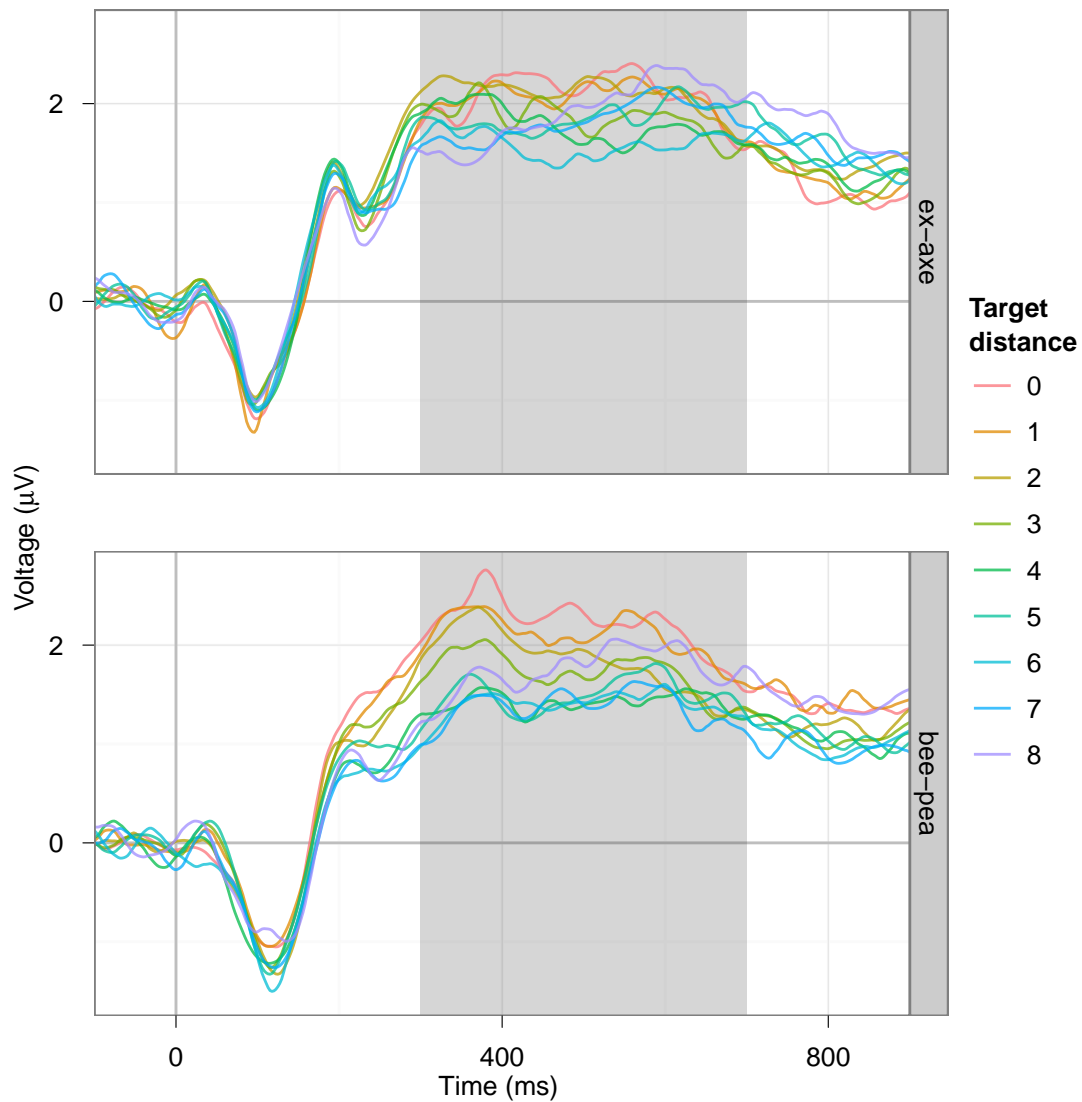
Figure 4.11: Experiment 4 results — ERP waveforms for parietal channels by target distance. The top panel shows ERPs for the *ex-axe* stimuli, and the bottom panel shows ERPs for the *bee-pea* stimuli. Target distance is computed from the step along the continuum and the target block (i.e., when *axe* is the target, step 9 along the *ex-axe* continuum corresponds to a target distance of 0, and step 1 along that continuum corresponds to a distance of 8). P3 amplitude decreases with distance from the target endpoint. Shaded areas indicate the time range used to compute mean P3 amplitude.
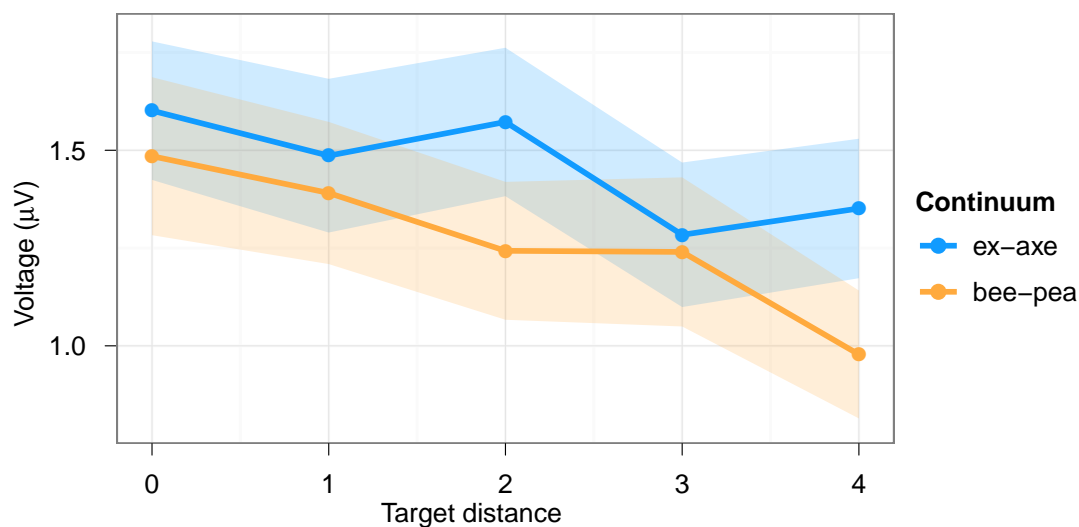
Figure 4.12: Experiment 4 results — P3 amplitude. Mean amplitude for the target trials as a function of distance from the target endpoint for the two stimulus continua. The top panel shows ERPs for the *ex-axe* stimuli, and the bottom panel shows ERPs for the *bee-pea* stimuli.

indicate steps within the target phonetic category. Because participants category boundaries varied, these distances contained different numbers of points. A few subjects with extreme category boundaries had distances more than 5 steps from their boundary; for most subjects, the target endpoint was 2-4 steps from their boundary.

Figure 4.13 shows grandaverage ERP waveforms and Figure 4.14 shows mean P3 amplitude as a function of distance from participants' category boundaries. A linear mixed-effects model with relative target distance and target (*bee* vs. *pea*) as fixed effects (r_{max}=0.068) showed a main effect of target distance (b=0.14, p_{MCMC}<0.001) with larger P3s for the steps closest to the target endpoint. There was a marginal effect of target (b=-0.25, p_{MCMC}=0.062) and the interaction was not significant (b=-0.08, p_{MCMC}=0.175).

This replicates the results of Toscano et al. (2010) showing that the P3 is sensitive to acoustic differences within a single phonetic category, even when subjects' responses are controlled for. Moreover, because there were sufficient "target" responses across the range

of VOT values, we can see that this effect follows a linear trend even across participants' category boundaries.

The same two analyses were also performed on the *ex-axe* data to assess whether the P3 also shows effects of within-category acoustic differences for these stimuli. As with the *bee-pea* stimuli, a linear mixed-effects model with target type entered as a fixed effect showed that P3 amplitude was significantly larger for the target (i.e., *ex* and *axe*) blocks (b=0.93, p_{MCMC}=0.002). Next, a model examining the target-block trials with target distance as a fixed effect showed a main effect (b=-0.042, p_{MCMC}=0.012) such that P3 amplitude decreased with increasing distance from the target endpoint.

Finally, target distance relative to participants' category boundaries was computed using the same procedure used for the *ex-axe* data. A linear mixed-effects model for the target-response trials was run with category boundary distance and target word (*ex* vs. *axe*) as fixed effects (r_{max}=-0.029). The model showed a main effect of boundary distance (b=0.08, p_{MCMC}=0.001), confirming that P3 amplitude shows effects of within-category acoustic differences. The effect of target (b=-0.01, p_{MCMC}=0.795) and interaction (b=0.04, p_{MCMC}=0.278) were not significant.[7]

4.3.3   Discussion

The results of this experiment are generally consistent with those of Toscano et al. (2010) showing that the N1 reflects continuous cue encoding. This was seen in the overall linear effect of continuum step for the *ex-axe* stimuli when listeners' responses were taken into account, and in the effect for the *bee-pea* stimuli in the overall dataset. However, the results were not completely unambiguous as the *bee-pea* stimuli only showed a marginal effect of step in the response-grouped dataset, and the *ex-axe* stimuli only showed an effect of F1 step when the data were grouped by response.

---

[7]Two additional participants were excluded from this analysis due to poor category boundary estimates obtained in the 2AFC task.
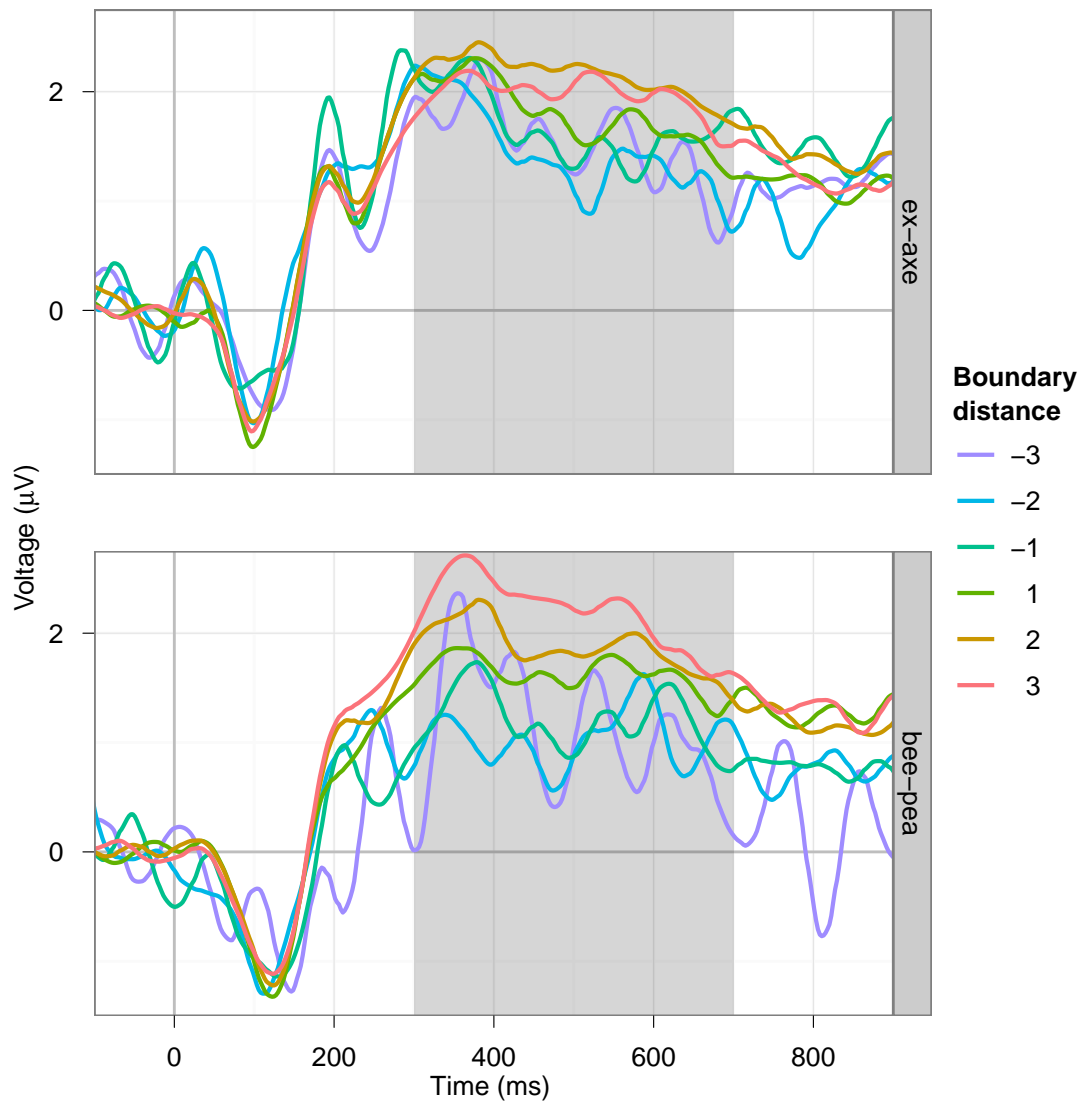
Figure 4.13: Experiment 4 results — ERP waveforms for parietal channels by category boundary distance. The top panel shows ERPs for the *ex-axe* stimuli, and the bottom panel shows ERPs for the *bee-pea* stimuli. The data consist of the trials on which listeners' made a "target" response. Positive step numbers indicate relative steps on the same side of the category boundary as the target endpoint, and the magnitude of the number is the number of steps from the listeners' category boundary for that continuum. Shaded areas represent the time range used to compute mean P3 amplitude.
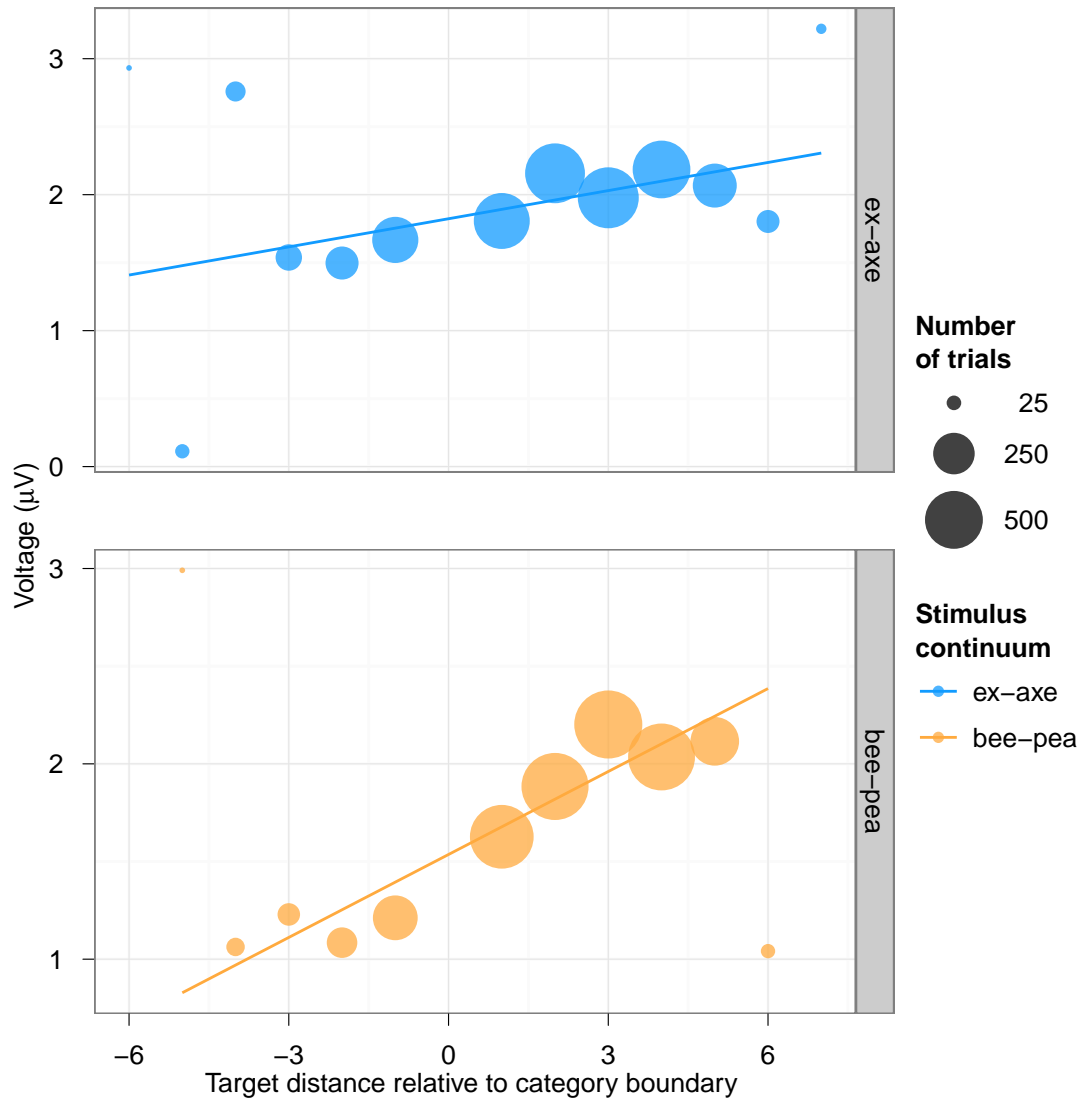
Figure 4.14: Experiment 4 results — P3 amplitude by category boundary distance. Mean amplitude for the target-response trials as a function of distance from listeners' category boundaries (positive values indicate steps closer to the target endpoint). The top panel shows ERPs for the *ex-axe* stimuli, and the bottom panel shows ERPs for the *bee-pea* stimuli. The size of the data point indicates the number of trials in that condition, and lines represent weighted linear models.

Although there was no indication of category-level influences on the N1 for the F1 differences in the *ex-axe* stimuli, the effects were smaller than those for VOT. There are several possible reasons for this. First, N1 amplitude was smaller overall for these stimuli. This fits with data showing that N1 responses are smaller for tones in this frequency range (Picton et al., 1978). Thus, it may have been more difficult to detect a linear trend in the N1 differences for theses stimuli. Second, unlike the stimuli in Experiment 1, the *ex-axe* stimuli used here had an ambiguous gender. Given that gender may affect how listeners' process differences in F1 (one of the proposed context compensation mechanisms) and the more variable N1 amplitudes observed for the ambiguously-gendered stimuli in Experiment 3, the lack of gender information here may have had an effect on listeners' N1 responses. However, the effect of F1 step on N1 amplitude seen in the response-grouped dataset does suggest that we can measure continuous encoding of F1 information.

In addition, the direction of the effect as a function of overall sound frequency varied for the two stimulus continua. For the VOT stimuli, lower VOTs, which have a lower over-all frequency, produced larger N1s. In contrast, stimuli with a lower F1, which also have a lower overall frequency, produced smaller N1s. This suggests that N1 to speech sounds may not simply reflect mean sound frequency, as it does for tones (Picton et al., 1978), and the re-lationship between the acoustic signal and N1 amplitude may be more complex for different speech sounds (e.g., we may be measuring two different frequency-based representations for the two cues).

Finally, the P3 results replicate those of Toscano et al. (2010) and extend them to show that within-category acoustic differences in vowels are maintained a post-perceptual stages as well. In addition, because listeners responses were more variable in this experi-ment, P3 amplitude could be examined across each continuum, demonstrating a linear ef-fect even across participant's category boundaries. This argues strongly against categorical models of speech perception.

Given this set of results, we can now ask whether preceding talker gender context affects listeners' categorization of ambiguously-gendered stimuli. This will be addressed in the next experiment.

4.4 Experiment 5: Effect of talker on vowel judgments

Extrinsic encoding approaches predict that listeners use preceding context information to compensate for talker-induced acoustic variability, either via cue-level (lateral) interactions or via feedback from more abstract representations about the talker. However, it is unclear whether listeners will use such information if it is in a *preceding* carrier phrase and target words have an ambiguous gender, since they may interpret the sentential context and target word as being spoken by two different talkers. While there is some evidence from work on nonspeech context information suggesting that listeners still use this type of extrinsic information in this situation (Lotto & Kluender, 1998; Holt, 2005), these effects are based on preceding information from tones rather than two different talkers.

To test this, I presented listeners with a 2AFC word identification task using the vowel stimuli from Experiment 3 in the context of a carrier phrase spoken by either a man or woman. Given the phonetic data on variation in /ɛ/ and /æ/, we would expect listeners to compensate for differences between men and women's voices such that they show more /æ/ (higher F1) responses in the context of a male talker (lower F1) and more /ɛ/ (lower F1) responses in the context of a female talker.

4.4.1 Methods

4.4.1.1 *Participants*

Eleven people participated in the experiment. Participant requirements, consent, and compensation procedures were the same as in Experiment 1. One participant was excluded from the analysis because of a problem with data collection.

4.4.1.2 Design

Listeners performed a 2AFC word identification task in which they heard the *ex-axe* stimuli from Experiment 3 in the context of either a male or female talker. Stimuli were presented in random order, and there were nine steps along the vowel continuum. Each stimulus was repeated 15 times for a total of 270 trials (9 vowel steps x 2 talker contexts x 15 repetitions). The experiment took approximately 30 minutes, completed in a single session.

4.4.1.3 Stimuli

The same *ex-axe* stimuli used in Experiment 3 were used here. The target words were spliced onto the end of carrier phrases ("On this trial, the word is...") spoken by the two talkers used for the original recordings. Recordings were made using the same using the same equipment used in Experiment 1, and the two talkers recorded the sentences with a similar prosodic structure, speaking rate, and with the word "did" at the end of the sentence to minimize coarticulation with the end of the carrier phrase. The highest-quality token from each talker was selected for the experiment, and the final word was removed from the sentence to create the carrier phrase used in the experiment.

The carrier phrases were equated for duration using the pitch synchronous overlap-add method (Moulines & Charpentier, 1990) by shortening the longer sentence (the token from the female talker) to match the duration of the shorter sentence. Next, they were low-pass filtered to 11.125 kHz using the same filter used with the target words (see Section 3.5.1.3). They were then equated for intensity and set to a level close to that of the target words. Finally, the target words were spliced onto the end of the carrier phrase. There was 150 ms of silence after the end of the carrier and before the onset of the target word.

4.4.1.4 Procedure

Participants were seated comfortably in front of a computer attached to a 19" CRT monitor in a sound-attenuated room. Participants responded by pressing one of two un-

marked buttons on the keyboard. During the experiment, the words "axe" and "ex" were displayed on the screen on the side corresponding to the button for that response. Response side was alternated between participants and the mapping between the words in the display and the keys on the keyboard was constant for each participant. On each trial, the two words appeared on the screen, followed the auditory stimulus 600 ms later. After making their response, the screen was cleared. Participants were offered a break every 45 trials.

4.4.2   Results

As in Experiment 2, listeners' responses reflected standard categorization functions, and they accurately identified both vowel endpoints (mean accuracy: 94.1%). There was also a shift in the mean number of /ɛ/-/æ/ responses as a function of talker context, such that there were more /æ/ responses in the context of a male talker than in the context of a female talker. This is consistent with the prediction from the phonetic data. Since men have lower formant values than women, listeners should shift their responses toward the /ɛ/ (lower F1) endpoint to compensate. Figure 4.15 shows listeners responses as a function of F1 step and talker gender.

The data were analyzed using a logit mixed-effects model with F1 and talker gender as fixed effects (r_{max}=-0.129). As expected, there was an effect of F1 step (b=0.83, z=27.36, p<0.001), demonstrating that listeners' categorized the ambiguously-gendered stimuli on the basis of F1. There was also an effect of talker gender (b=-0.45, z=-3.94, p<0.001), indicating that the talker context had an effect on listeners' responses. The size of the effect appears similar at each F1 step, though it is difficult to assess whether or not this is actually the case since the data are binomial. The talker gender x F1 step interaction was not significant (b=0.05, z=0.90, p=0.369).
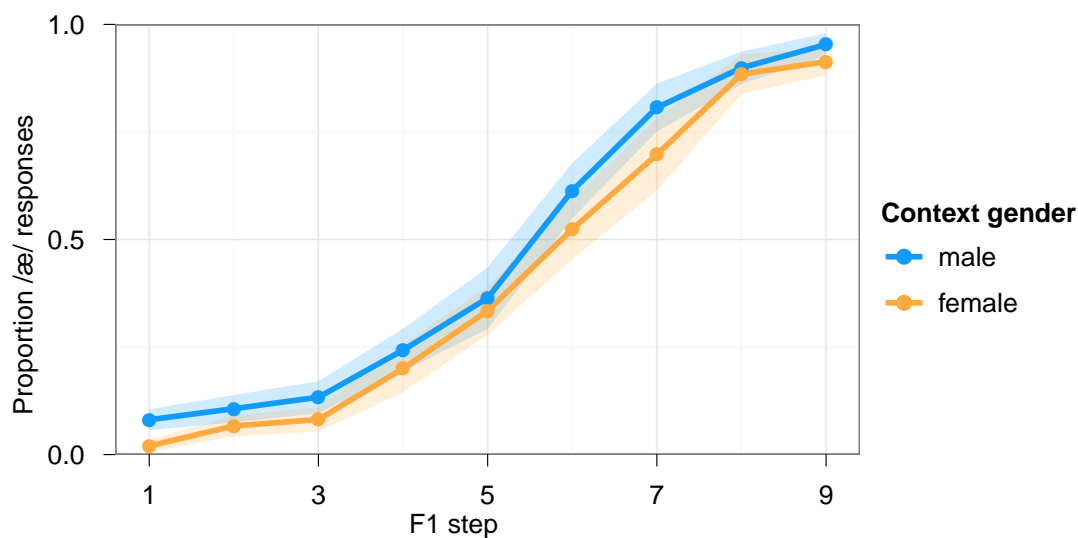
Figure 4.15: Experiment 5 results — categorization responses. Proportion of /æ/ responses as a function of F1 step and gender of the talker in the preceding carrier phrase. Listeners' made more /æ/ responses in the context of a male talker than in the context of a female talker.

4.4.3 Discussion

These results demonstrate that preceding talker gender influences listeners' vowel judgments, even when the vowel sound is presented with an ambiguously-gendered voice. This fits with previous work showing that talker gender can be signaled by auditory information within the target word itself (Strand & Johnson, 1996), as well as by information outside the phonetic segment that indicates gender (e.g., visual information ; Johnson et al., 1999). In addition, it extends the results of (Ladefoged & Broadbent, 1957) showing that talker variability in the preceding sentence can affect vowel judgments, by showing that this is the case even when talker differences are removed from the target word.

The next experiment examines the locus of this effect using the ERP paradigm from Experiment 4 and the carrier phrases from the present experiment.

4.5   Experiment 6: Effect of talker on cue encoding

The final experiment in this series examines the effect of talker context. Listeners heard the carrier phrases from Experiment 5 spliced onto the target words from Experiment 4, and the effect of talker on N1 and P3 amplitude was examined.

If listeners compensate for talker differences on the basis of preceding talker context by encoding relative cue-values, as predicted by extrinsic approaches, we would expect differences in N1 amplitude as a function of talker gender. Specifically, for F1 differences, we would expect larger N1s in the context of a male talker if listeners recode F1 relative to the preceding talker information. That is, because men have lower F1-values overall, listeners would compensate by recoding F1 values as higher than they actually are.

Similarly, because VOT also varies as a function of talker, extrinsic approaches predict differences in N1 amplitude to stimuli varying in VOT as a function of talker gender. Because women have longer VOTs than men, and because N1 amplitude *decreases* with increasing VOT, we would expect listeners to show larger N1s in the context of female talkers for VOT distinctions if they encode VOT values relative to talker gender. Thus, relative encoding approaches to context compensation predict that effects of talker gender for the two stimulus continua will occur in opposite directions.

There are other ways that talker gender could affect N1 amplitude that would not reflect relative cue encoding processes. For example, there could be an overall difference in N1 amplitude for one talker gender than the other. One difference between male and female talkers that could drive such an effect is differences in speech intelligibility between the two groups. Recall that men have lower intelligibility scores than women (Bradlow et al., 1995). Because of this, listeners may show enhanced attention to the speech signal in the context of a male talker in order to extract as much information as possible. Since increased attention leads to enhancement of the N1 (Hansen & Hillyard, 1980), we would expect a larger N1 in the context of a male talker for both stimulus continua.

Thus, together, the two continua allow us to test this hypothesis, along with the different context compensation mechanisms. If both continua show the same effect of talker gender, it would suggest there is a general effect of talker gender on how listeners process cues in the target word. In contrast, if N1 amplitude is larger in the context of a male talker for the *ex-axe* continuum and smaller for the *bee-pea* continuum, it would suggest that listeners encode cues relative to context, supporting extrinsic encoding proposals. If context has no effect on N1 amplitude, it would be consistent with raw-cue encoding accounts.

4.5.1   Methods

4.5.1.1   *Participants*

Twenty-four people participated in the experiment. Participant recruitment, consent, and compensation procedures were the same as the previous experiments, and participants met the same language, hearing, and vision criteria as in previous experiments. Three participants were excluded for having less than 50% correct responses on at least one endpoint for both stimulus continua, and three additional participants were excluded from the *ex-axe* analyses for having less than 50% correct responses at one of the endpoints along that continuum.

4.5.1.2   *Design*

Participants performed a target detection task, similar to Experiment 4. Stimuli consisted of the sentences from Experiment 5, with the duration of the carrier phrase modified to vary over a 312 ms range in six steps. This reduces overlap from ERP components associated with the carrier phrase on the ERP components to the target word, and it allows us to use Adjar to remove additional overlap (see Section 3.2 for details of the Adjar procedure). Each stimulus was repeated once at each carrier duration in the four target blocks, for a total of 864 trials.

As in Experiment 4, participants also performed a 2AFC categorization task (with

five repetitions of each stimulus) after the ERP session to obtain estimates of their category boundaries. The experiment took approximately two hours, completed in a single session.

4.5.1.3 *Stimuli*

As in Experiment 1, the overlap-add method was used to vary the duration of the carrier phrase used in Experiment 5, creating different ISIs between the onset of the sentence and onset of the target word. ISIs varied in six steps from 1836 to 2148 ms.

4.5.1.4 *Procedure, EEG recording, and data processing*

The experiment setup was the same as the previous ERP experiments, and participants performed the same target detection task used in Experiment 4. Participants responded with whether the word they heard was the target or not, and each word (*axe*, *ex*, *had*, and *head*) served as the target in a separate block. EEG recording and data processing procedures were the same as those used in Experiments 1-4.

4.5.2   Results

4.5.2.1   *Behavioral responses*

The participants included in the final dataset correctly categorized the stimulus endpoints in each target block (mean accuracy: 97.3%). As in previous experiments, responses varied with target block such that listeners were more likely to indicate that a stimulus belonged to the target category.

Figure 4.16 shows listeners' responses as a function of step and talker gender for the carrier phrase. As in Experiment 4, there were more *axe* responses in the context of a male talker than in the context of a female talker. In addition, there were more /p/ responses in the context of a male talker. This fits with phonetic data showing that women have longer VOTs than men and suggests that listeners compensated for differences in men and women's VOT values when making voicing judgments.

These observations were evaluated statistically using logit mixed-effects models for each stimulus continuum. For the *bee-pea* stimuli, VOT step, target (*bee* vs. *pea*), and talker gender were entered as fixed effects (r_{max}=0.385). As expected, the model showed a main effect of VOT step (b=1.56, z=30.51, p<0.001). There was also a main effect of target (b=1.57, z=12.23, p<0.001), which fits with the observation from previous experiments that listeners' made more target than nontarget responses. In addition, there was a main effect of talker gender (b=0.55, z=4.42, p<0.001), with listeners making more /p/ responses in the context of a male talker than a female talker. This demonstrates that listeners compensate for VOT differences between men and women. Finally, there was a significant VOT step x gender interaction (b=-0.26, z=-2.87, p=0.004), indicating that the VOT categorization function was steeper in the context of a male talker than in the context of a female talker.

A corresponding model was run for the *ex-axe* continuum (r_{max}=0.348). There was a main effect of F1 step (b=.17, z=32.94, p<0.001), as well as a main effect of target (b=1.35, z=12.16, p<0.001), with a similar pattern of results to the *bee-pea* stimuli (i.e., more responses for the endpoint that served as the target). There was also a main effect of gender (b=0.40, z=3.64, p<0.001), replicating the results of Experiment 4 and showing that listeners compensated for differences in gender in their vowel categorization responses. There was a marginal F1 step x target interaction (b=0.13, z=1.89, p=0.059) and a significant F1 step x target x gender interaction (b=0.29, z=2.13, p=0.033), indicating that the slope of listeners' categorization functions varied depending on the specific combination of talker gender and target. Other interactions were nonsignificant.

Listeners responses during the 2AFC task run after the ERP session also showed normal categorization functions. As in Experiment 3, the data were fit to four-parameter logistic curves to obtain each participant's category boundary. The mean category boundary for the *bee-pea* continuum was at step 5.2, and the mean boundary for the *ex-axe* continuum was step 5.5 (with steps ranging from 1 to 9).
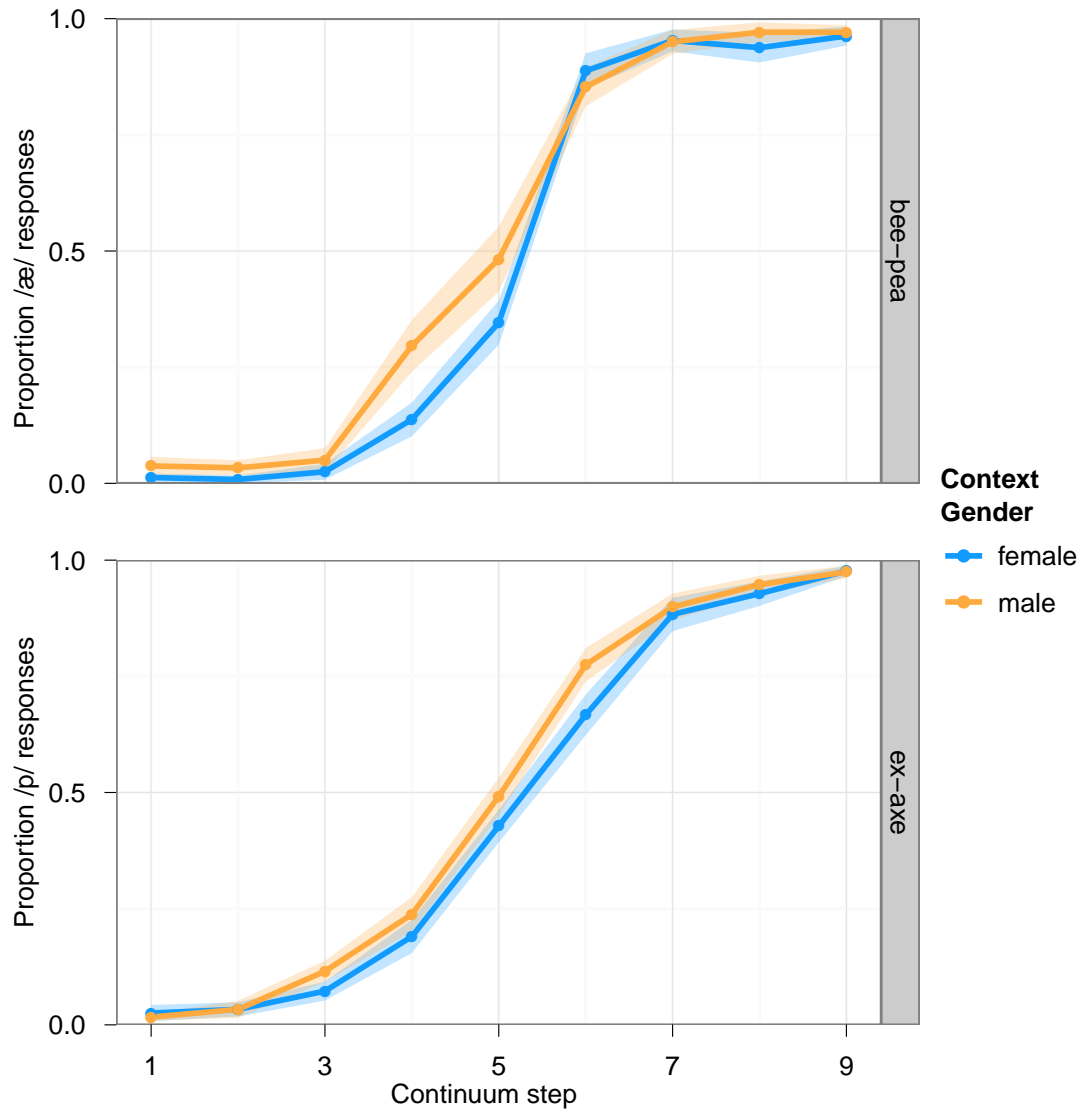
Figure 4.16: Experiment 6 results — behavioral responses during target detection task. Categorization responses as a function of continuum step and talker gender in the preceding carrier phrase. Listeners made more /p/ responses in the context of a male talker for the *bee-pea* stimuli (top panel) and more /æ/ responses for the *ex-axe* stimuli (bottom panel).

4.5.2.2   N1 amplitude

Before examining N1 responses, the Adjar procedure was used to remove overlap between ERPs to the carrier phrase and those to the target word. Recall that the ISI between the onset of the carrier phrase and the onset of the target word varied over a 312-ms interval. Examination of ERPs time-locked to the onset of the carrier phrase did not reveal any systematic variation in the time period prior to the onset of the target word. This isn't surprising since the carrier phrases were relatively long (1836 to 2148 ms). However, there did appear to be overlap from the carrier in the pre-stimulus baseline period for the target word. This appeared to be due to ERPs produced by the offset of periodic energy in the carrier phrase, which occurred 140 to 160 ms before the onset of the target word. Indeed, time-locking the ERPs to that event revealed that the overlap seen in the pre-stimulus period appeared to be related to it. Although the ISI time range between the offset of the carrier and the onset of the target word was relatively small (20 ms), the Adjar procedure was applied, and it converged on a stable estimate of the overlap. Figure 4.17 shows the ERP waveforms for the two talker gender conditions time-locked to the onset of the target word before and after Adjar. As the figures shows, the amount of overlap in the pre-stimulus period was reduced by the Adjar procedure.

N1 responses as a function of continuum step and context gender were then examined. Figure 4.18 shows grandaverage ERP waveforms for the average of the frontal channels as a function of step, and Figure 4.19 shows waveforms as a function of talker gender during the carrier phrase. Overall N1 latency was later than in Experiment 4, which presented these stimuli in isolation. As such, the time window used to compute mean N1 amplitude was adjusted to capture the majority of the component for each stimulus continuum. For the *bee-pea* stimuli, mean N1 was computed from 130 to 170 ms post-stimulus, and for the *ex-axe* stimuli, mean N1 was computed from 115 to 165 ms post-stimulus. Figure 4.20 shows mean N1 amplitude for the two stimulus continua.
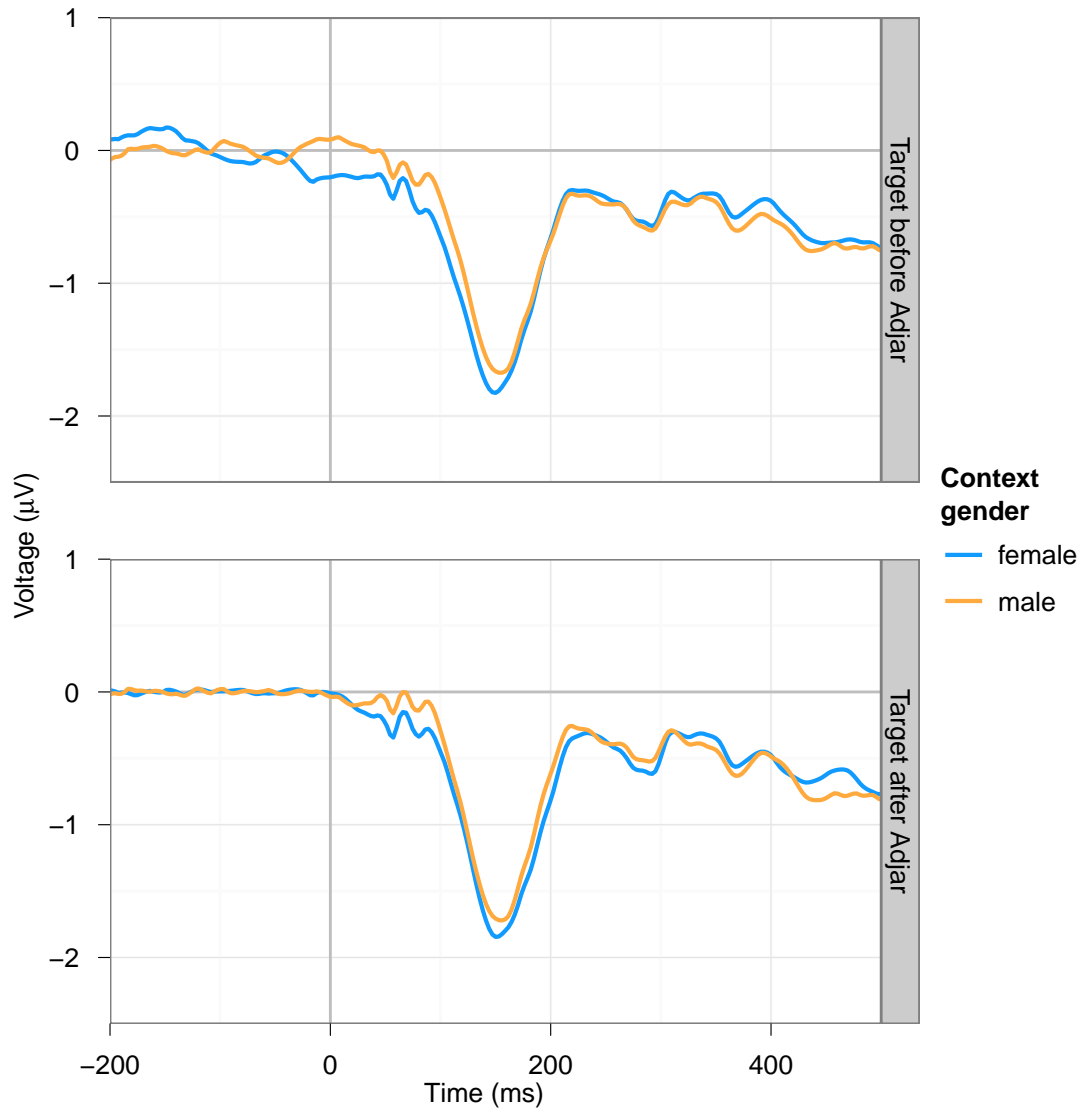
Figure 4.17: Experiment 6 results — effectiveness of Adjar procedure. Grandaverage ERP waveforms for the average of the three frontal channels time-locked to the onset of the target word before (top panel) and after (bottom panel) application of the Adjar procedure. For both talker gender conditions, the amount of overlap in waveforms was reduced after the Adjar technique was used.

The same set of analyses used in Experiment 3 was used here, with the addition of talker gender as an additional factor. First, to evaluate effects of continuum step and talker gender on N1 responses to the *bee-pea* stimuli, a linear mixed-effects model was run with VOT step, talker gender, and target type (*bee/pea* vs. *ex/axe*) entered as fixed effects (r_{max}=0). There was a main effect of talker gender (b=-0.29, p_{MCMC}<0.001) with larger N1s in the context of the female talker. This fits with the prediction of the relative encoding hypothesis which suggests that listeners' compensate for VOT differences between talkers by treating VOTs as shorter than they actually are (hence, a larger N1) in the context of a female talker (since they have longer VOTs), and as longer than they are (smaller N1) in the context of a male talker (since they have shorter VOTs).

Despite this result, however, a main effect of VOT step was not observed (b=0.016, p_{MCMC}=0.198). This could be due to overlap from later ERP components, especially given the longer latency N1s observed in this experiment, or it could reflect a difference in listeners' processing of VOT in running speech. Finally, there was a marginal effect of target type (b=-0.12, p_{MCMC}=0.060). All interactions were nonsignificant (VOT x target type: b=-0.023, p_{MCMC}0.357; VOT x gender: b=0.0033, p_{MCMC}=0.911); target type x gender: b=0.036, p_{MCMC}=0.772; VOT x target type x gender: b=0.047, p_{MCMC}=0.352).

Next, target-response trials were analyzed as a function of VOT step, talker gender, and target (*bee* vs. *pea*). Figure 4.21 shows mean N1 amplitude for these trials as a function of continuum step, and Figure 4.22 shows mean N1 amplitude as a function of talker gender. In both cases, data points are weighted by the number of trials in each condition. A linear mixed-effects model with those factors entered as fixed effects (r_{max}=-0.710) found a main effect of talker gender (b=-0.43, p_{MCMC}<0.001), supporting the result seen in the overall data. There was a marginal effect of VOT step (b=0.046, p_{MCMC}=0.052) with N1 amplitude decreasing with increasing VOT. This suggests that listeners' may be encoding continuous VOT differences in the context of the different talkers. Neither the main effect of target
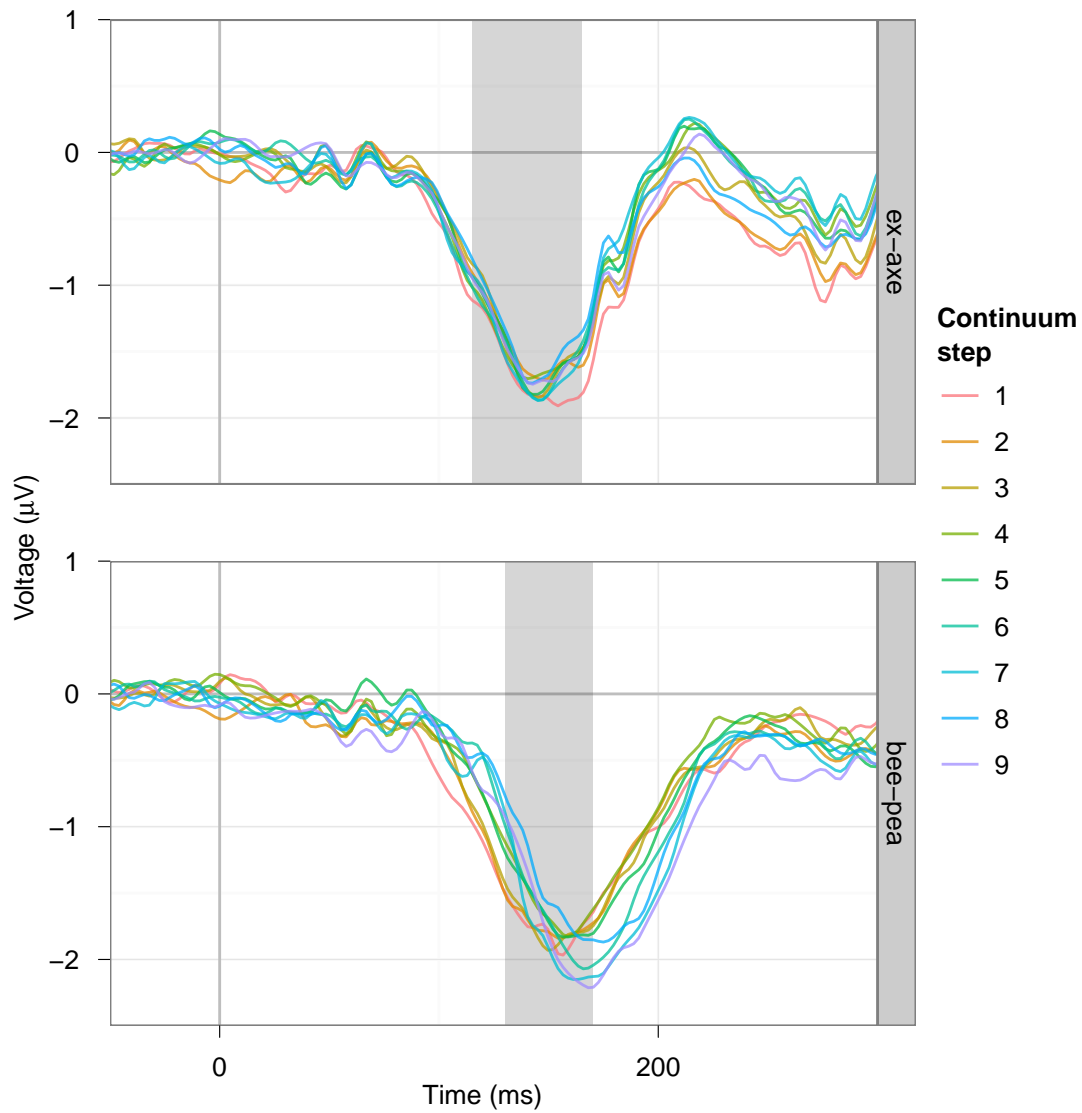
Figure 4.18: Experiment 6 results — ERP waveforms for frontal channels by continuum step. The top panel shows ERPs for *ex-axe* stimuli, and the bottom panel shows ERPs for the *bee-pea* stimuli. Shaded areas indicate time ranges used to compute mean N1 amplitude.
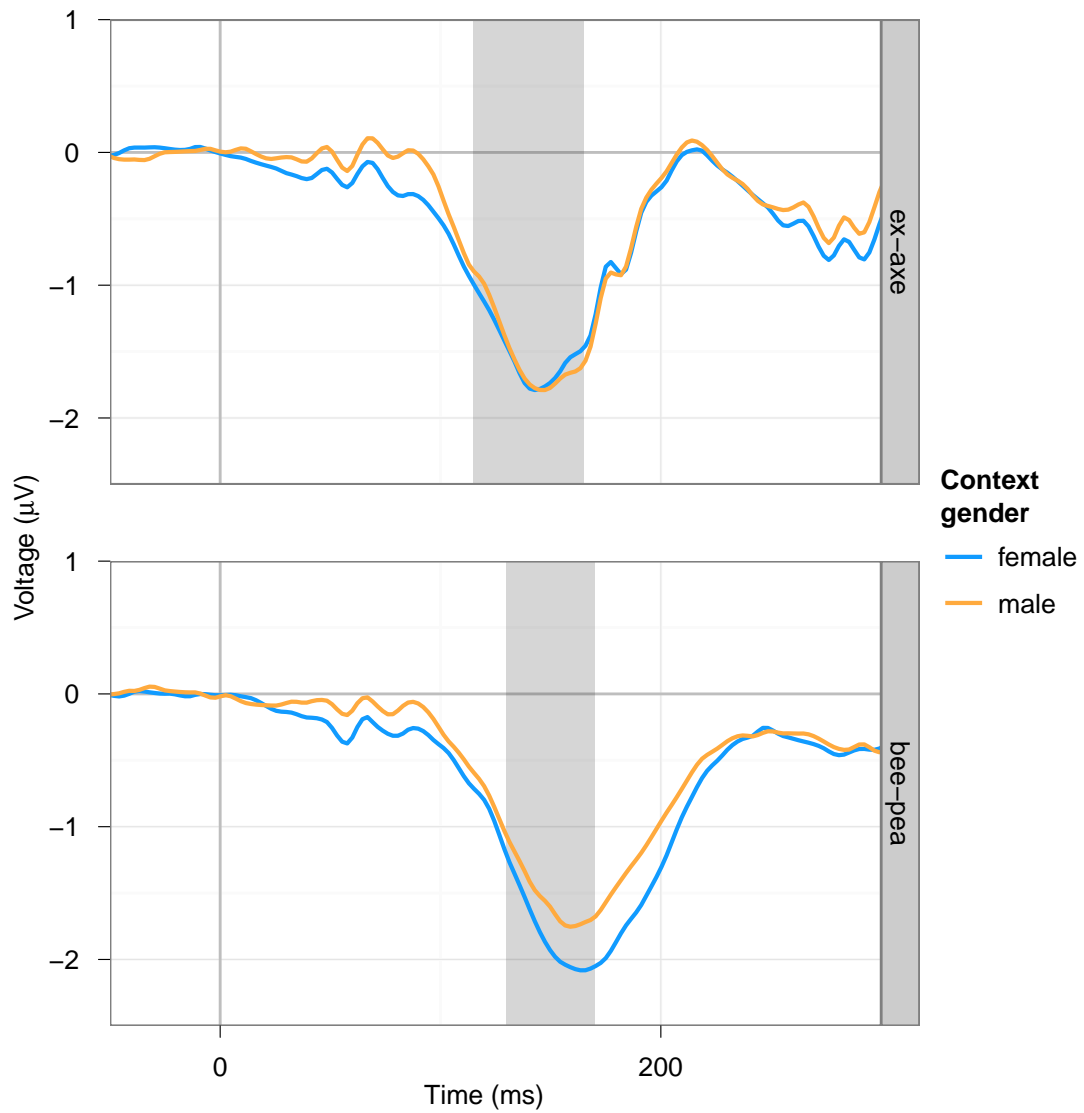
Figure 4.19: Experiment 6 results — ERP waveforms for frontal channels by talker gender. The top panel shows ERPs for the *ex-axe* stimuli, and the bottom panel shows ERPs for the *bee-pea* stimuli. Shaded areas indicate time range used for N1 amplitude.
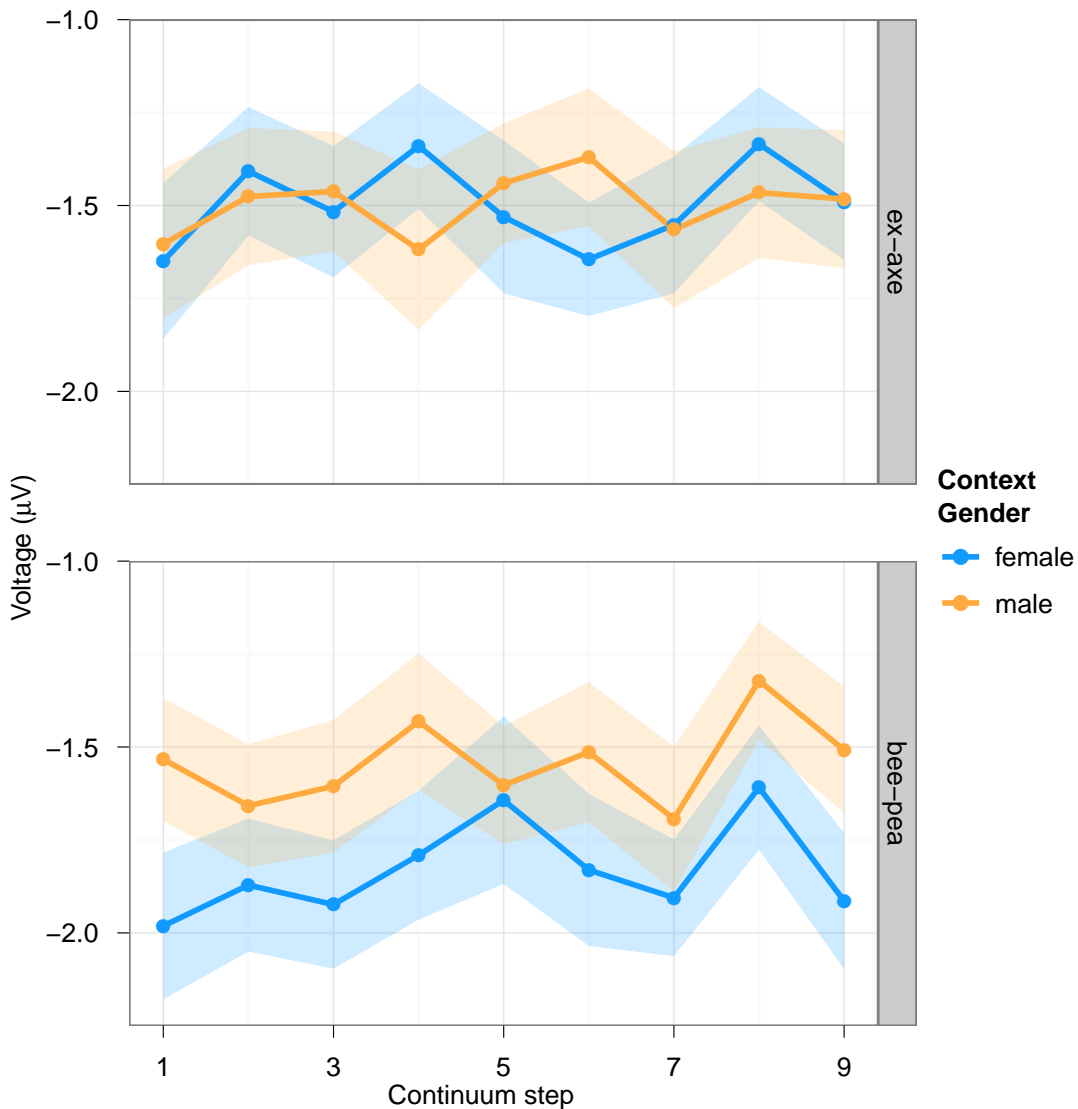
Figure 4.20: Experiment 6 results — N1 amplitude. Mean amplitude as a function of continuum step and talker gender. The top panel shows N1s for the *ex-axe* stimuli, and the bottom panels shows N1s for the *bee-pea* stimuli. Shaded areas indicate standard error.

nor any of the interactions were significant (target: b=0.10, p_{MCMC}=0.390; VOT x target: b=-0.037, p_{MCMC}=0.434; VOT x gender: b=-0.012, p_{MCMC}=0.809; target x gender: b=0.17, p_{MCMC}=0.473; VOT x target x gender: b=0.067, p_{MCMC}=0.475).

Overall, these results are consistent with the proposal that listeners' encode cues relative to context. However, it may be that the N1 is simply larger in the context of a female talker for some other reason. If that is the case, we would expect the same pattern of results for the *ex-axe* stimuli. In contrast, if the gender effect observed for the *bee-pea* stimuli reflects relative cue encoding, we would expect the opposite pattern of results: larger N1s (reflecting higher F1 frequencies) in the context of a male talker than in the context of a female talker.

To test this, mean N1 amplitude for the *ex-axe* stimuli was examined using the same two analyses. First, a linear mixed-effects model with F1 step, talker gender, and target type was run on the overall dataset (r_{max}=0). This showed a main effect of target type (b=0.38, p_{MCMC}<0.001), with larger N1s for the *bee/pea* target blocks than for the *ex/axe* target blocks (consistent with the results of Experiment 4), but no effect of F1 (b=0.0086, p_{MCMC}=0.480), gender (b=0.0018, p_{MCMC}=0.982), or any interactions (F1 x target type: b=-0.0045, p_{MCMC}=0.859; F1 x gender: b=-0.0013, p_{MCMC}=0.954; target type x gender: b=-0.022, p_{MCMC}=0.850; F1 x target type x gender: b=-0.021, p_{MCMC}=0.666).

Next, N1 amplitude was analyzed for the target-response trials, since this showed a more robust effect of F1 in Experiment 4. The model included F1 step, target (*ex* vs. *axe*), and gender as fixed effects (r_{max}=-0.691). There was a significant effect of target (b=0.26, p_{MCMC}=0.022), but no other significant effects (F1: b=0.0001, p_{MCMC}=0.989; gender: b=-0.030, p_{MCMC}=0.796; F1 x target: b=-0.033, p_{MCMC}=0.503; target x gender: b=-0.0078, p_{MCMC}=0.956; F1 x target x gender: b=-0.026, p_{MCMC}=0.786).

Although there were no effects of talker gender, the lack of an effect in these analyses suggests that the effect observed for the *bee-pea* stimuli not was purely driven by an over-
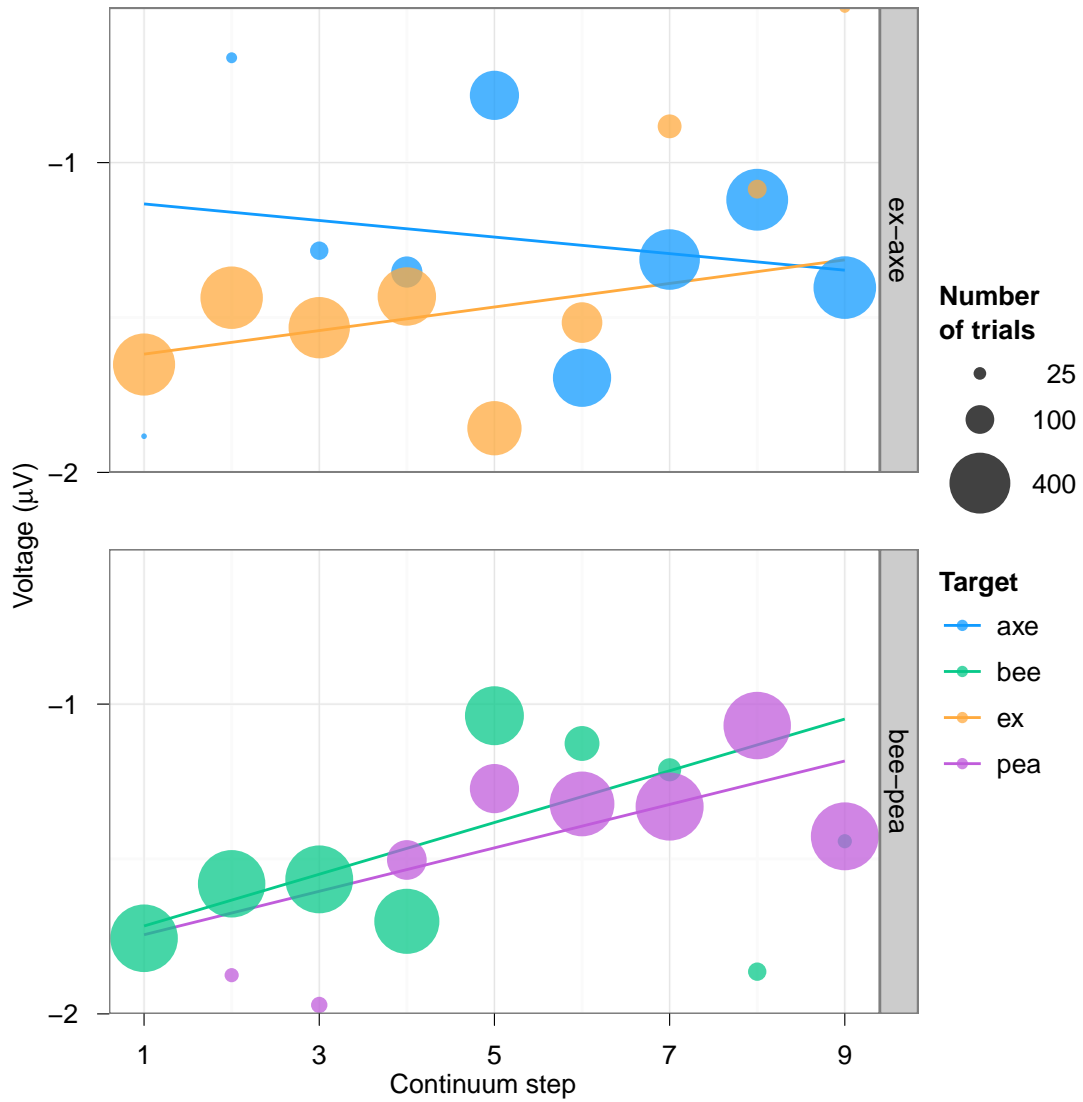
Figure 4.21: Experiment 6 results — N1 amplitude for target-response trials grouped by continuum step. Mean amplitude for the target-response trials as a function of continuum step and target word. The top panel shows ERPs for the *ex-axe* stimuli, and the bottom panel shows ERPs for the *bee-pea* stimuli. The size of the data points indicates the number of trials in each condition, and lines represent weighted linear models. Data from each condition are reflected in the models, though one data point (step 1 along the *bee-pea* continuum for the *pea*-target trials) is not plotted because it fell outside the range of values of the other points.
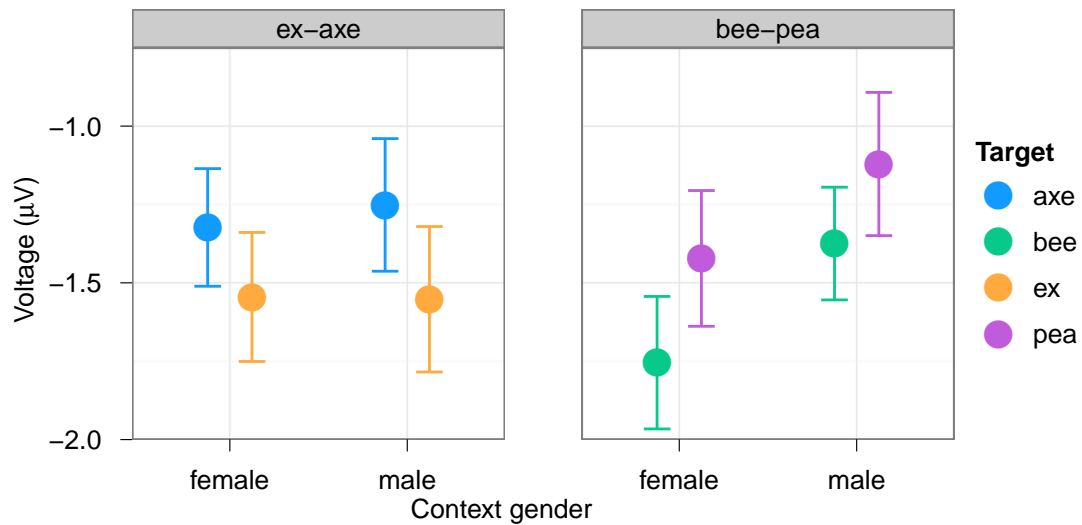
Figure 4.22: Experiment 6 results — N1 amplitude for target-response trials grouped by talker gender. Mean amplitude for the target-response trials as a function of preceding talker gender and target word. The left panel shows ERPs for the *ex-axe* data, and the right panel shows ERPs for the *bee-pea* data. Error bars indicate standard error.

all effect of talker gender caused by differences between listeners' processing of men and women's voices. Moreover, if there is an overall effect of talker gender for both continua, such that N1s are larger in the context of a female talker, any context compensation effects for the F1 stimuli may have been cancelled out and compensation effects in the VOT stimuli may have been amplified. Indeed, this fits the pattern of results seen here: a significant effect in the predicted direction for the VOT stimuli and a null effect for the F1 stimuli.

4.5.2.3   P3 amplitude

To see whether effects of talker context and gradiency along the acoustic cue continua could be observed at post-perceptual processing stages, mean P3 amplitude was analyzed similarly to Experiment 4. Figure 4.23 shows grandaverage waveforms for the parietal channels as a function of trial type (target vs. nontarget), showing that a large P3 is observed for the target trials. Figure 4.24 shows ERPs for the target trials as a function of

target distance, showing that P3 amplitude decreases with distance from the target end-point, replicating the results of previous experiments.

To examine the effect of talker context, talker gender was coded as a function of whether it provides information that is either more or less consistent with the target end-point. For example, in the context of a female talker, listeners produce more "bee" responses on average (compensating for longer VOT values). Thus, for the female talker context, we would expect larger P3s when *bee* is the target word, since, on average, perceived VOT values are shorter or listeners are biased to make "bee" responses. Conversely, in the context of a male talker, we would expect larger P3s when *pea* is the target. These two combinations of conditions were coded as "consistent". The opposite combinations (male talker, *bee* target; female talker, *pea* target) were coded as "inconsistent". Figure 4.25 shows grandaverage ERP waveforms for target trials as a function of whether the preceding context was consistent or inconsistent with the target endpoint.

Mean P3 amplitude was computed for the average of the three parietal channels from 300 to 700 ms post-stimulus. Figure 4.26 shows mean P3 amplitude as a function of target distance and context for each continuum.

A linear mixed-effects model with target type as a fixed effect was run on the *bee-pea* stimuli, and found a main effect (b=-0.85, p_{MCMC}<0.001) with larger P3s for the target (*bee* and *pea*) blocks than for the nontarget (*ex* and *axe*) blocks. Next, a model examining P3 amplitude on the target trials with target distance and context consistency as fixed effects was run (r_{max}=0). There was a main effect of target distance (b=-0.14, p_{MCMC}<0.001) with P3 amplitude decreasing with distance from the target endpoint. Neither the effect of context consistency nor the interaction were significant (context: b=0.047, p_{MCMC}=0.593; VOT x context: b=-0.019, p_{MCMC}=0.601).

Next, P3 amplitude was analyzed as a function of distance from each participants' VOT boundary using the same procedure used in Experiment 4. Figures 4.27 and 4.28 show
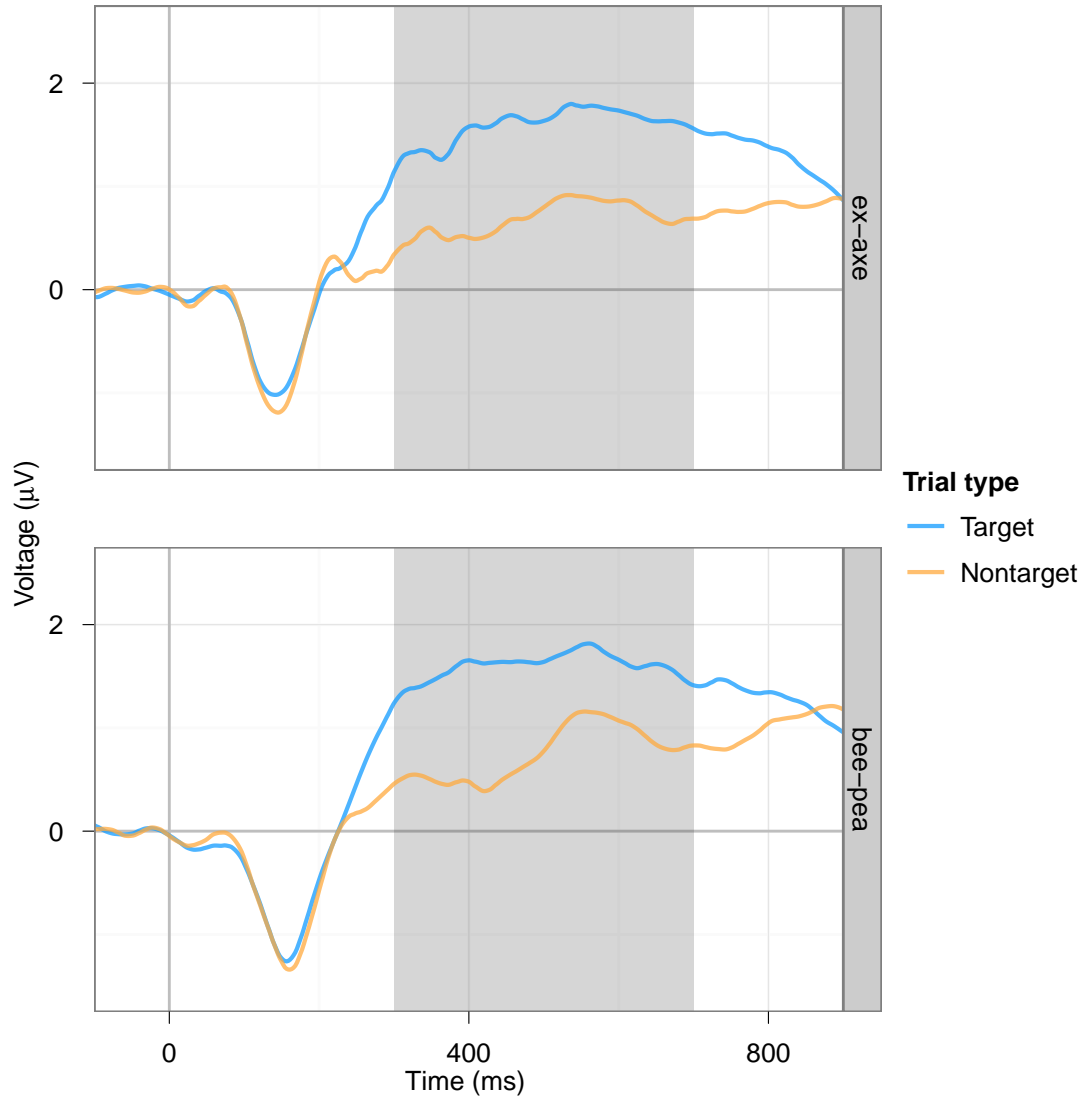
Figure 4.23: Experiment 6 results — ERP waveforms for parietal channels by trial type. For the *ex-axe* stimuli (top panel), the *ex* and *axe* target blocks correspond to the "target" trials, and the *bee* and pea target blocks correspond to the "nontarget" trials. For the *bee-pea* stimuli (bottom panel), *ex* and *axe* blocks are "nontarget" trials, and *bee* and *pea* blocks are "target" trials. For both continua, larger P3s were found for the target trials. Shaded areas indicate time range used to compute mean P3 amplitude.
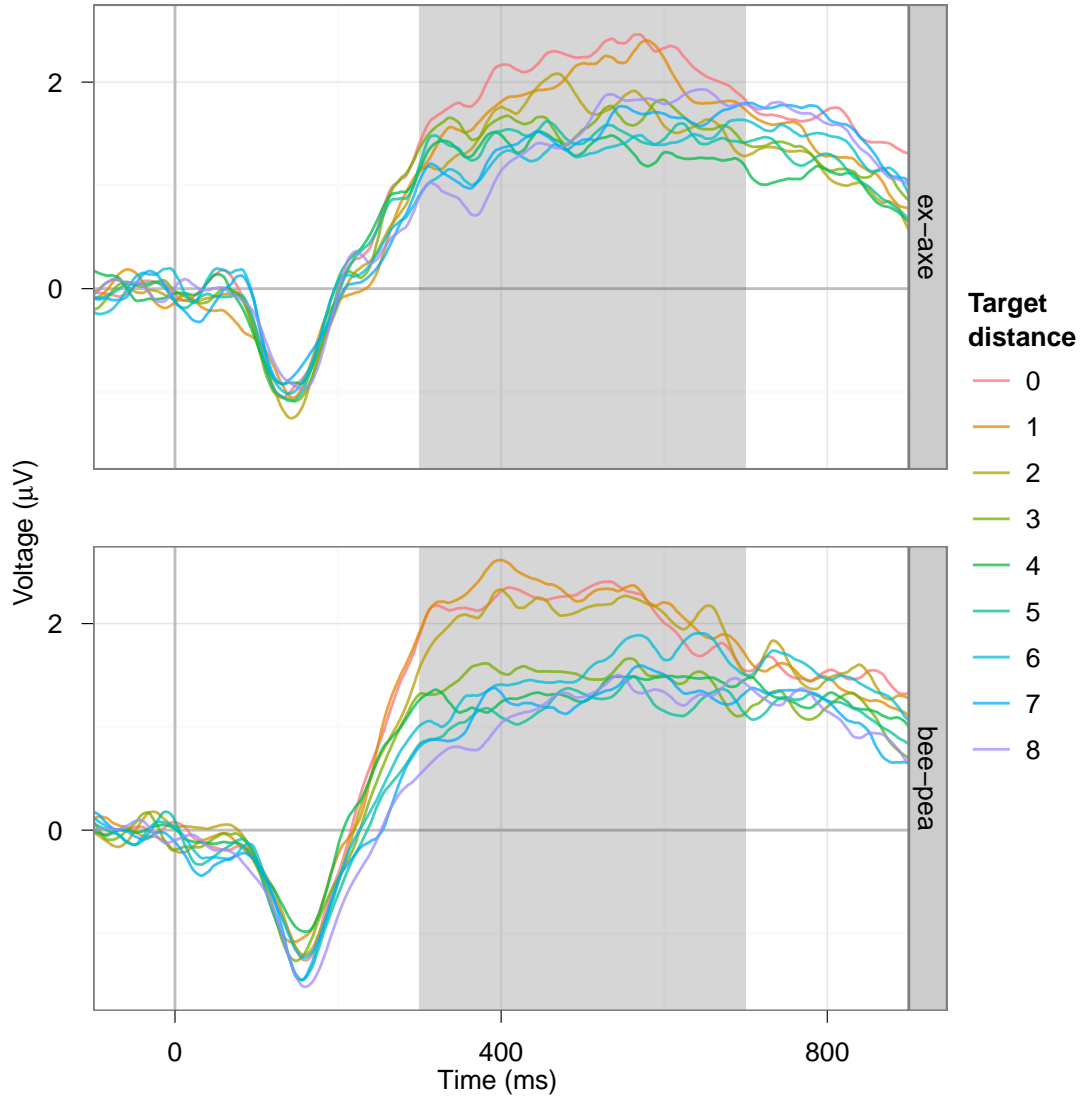
Figure 4.24: Experiment 6 results — ERP waveforms for parietal channels by target distance. Target distance is computed in terms of the number of continuum steps away from the relevant target endpoint (e.g., step 9 along the *ex-axe* continuum corresponds to a target distance of 0 when *axe* is the target). P3 amplitude decreased with distance from the target endpoint. Shaded areas indicate time range used for P3 amplitude.
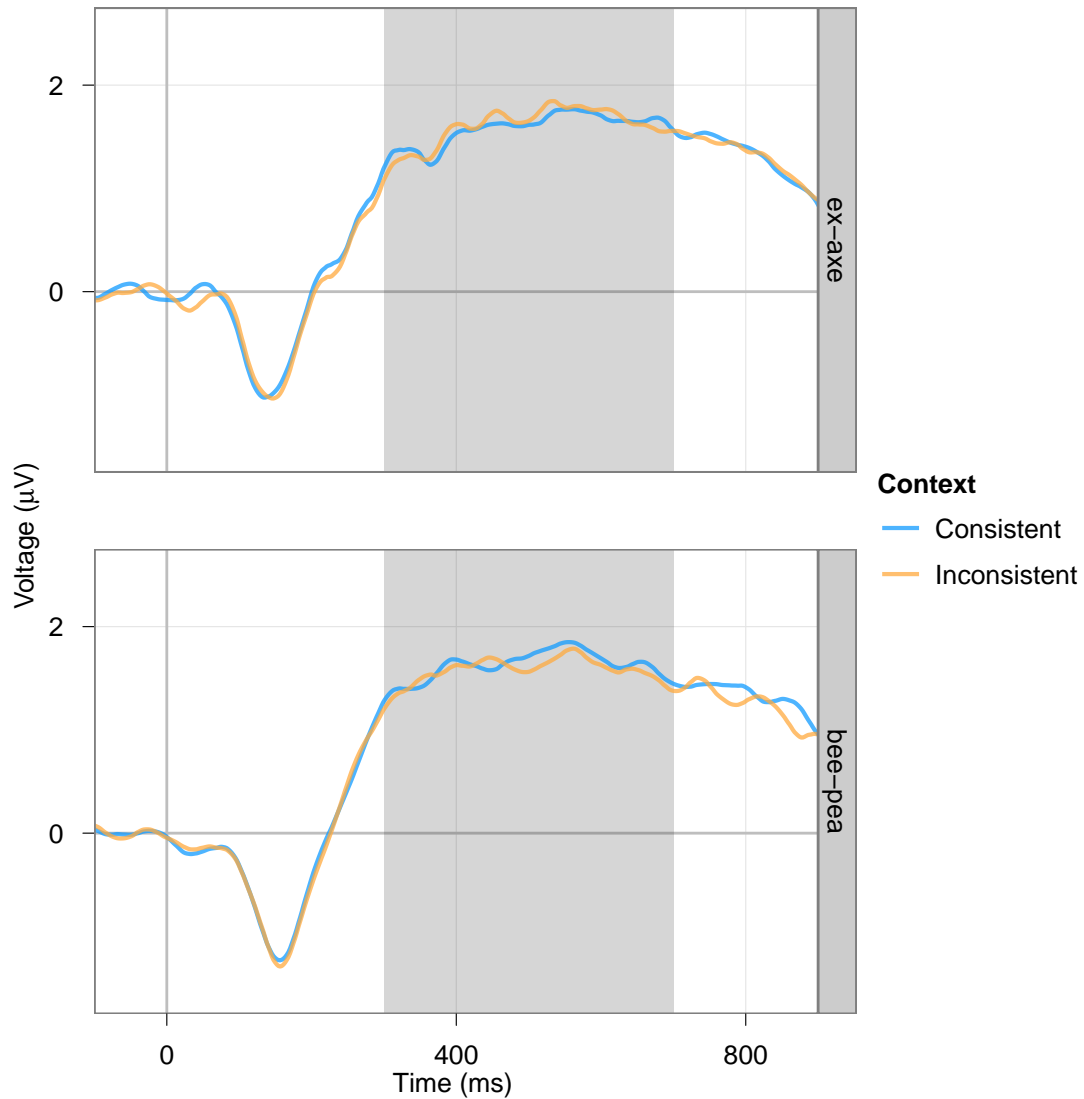
Figure 4.25: Experiment 6 results — ERP waveforms for parietal channels by context. Preceding context is coded based on whether it is consistent with the target endpoint. For example, in the context of a male talker, listeners make more "axe" responses. Thus, P3 responses for the *axe* target block are expected to be larger in the context of a male talker than a female talker and are coded here as "consistent". Shaded areas indicate time range used for P3 amplitude.
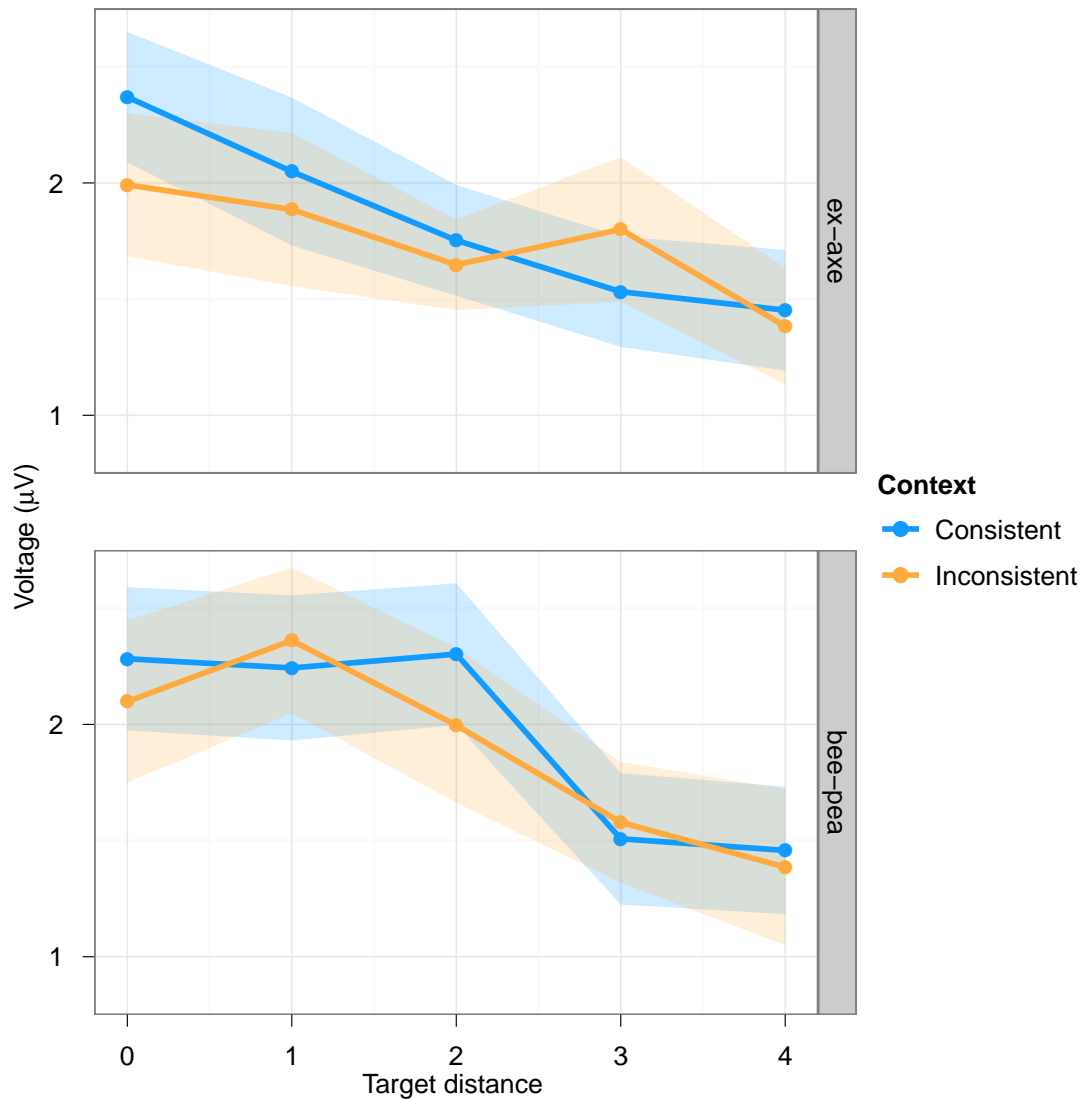
Figure 4.26: Experiment 6 results — P3 amplitude by target distance and context. Mean amplitude as a function of stimulus continuum, distance from the target endpoint, and whether the preceding context was consistent with the target endpoint. The top panel shows ERPs for the *ex-axe* stimuli, and the bottom panel shows ERPs for the *bee-pea* stimuli.

ERP waveforms for the target-response trials as a function of distance from the participants' category boundary and preceding context; Figure 4.29 shows mean P3 amplitudes for these data. A linear mixed-effects model with boundary distance, context consistency, and target (*bee* vs. *pea*) as fixed effects (r_{max}=-0.060) found a main effect of boundary distance (b=-0.17, p_{MCMC}<0.001), but no other significant effects (target: b=0.073, p_{MCMC}=0.543; context: b=-0.14, p_{MCMC}=0.199; distance x target: b=0.0005, p_{MCMC}=0.978; distance x context: b=0.045, p_{MCMC}=0.467; target x context: b=0.069, p_{MCMC}=0.782; distance x target x context: b=0.14, p_{MCMC}=0.260). Thus, although an effect of talker gender was found for listeners' overt responses and in the N1 data, this effect was not seen in the P3.

The same two analyses were used to examine the *ex-axe* data. As expected, a linear mixed-effects model with target type as a fixed effect was significant (b=-0.91, p_{MCMC}=0.004) with larger P3s for the target than for the nontarget conditions. P3 amplitude for the target trials was examined using a model with target distance and context consistency as fixed effects (r_{max}=0). There was a main effect of target distance (b=-0.076, p_{MCMC}<0.001) with P3 amplitude decreasing with distance from the target endpoint. The main effect of context was not significant (b=-0.024, p_{MCMC}=0.778), but there was a target distance x context interaction (b=-0.074, p_{MCMC}=0.026). A follow-up analysis found that the effect of target distance was significant when the context was consistent with the target (b=-0.11, p_{MCMC}<0.001), but not when it was inconsistent (b=-0.040, p_{MCMC}=0.108). Thus, talker gender had different effects depending on distance from the target.

As in the earlier P3 analyses, steps along the F1 continuum were coded relative to each participant's F1 boundary, and a linear mixed-effects model with boundary distance, target (*ex* vs. *axe*, and context consistency was run for the target-response trials (r_{max}= -0.166). The model showed a main effect of boundary distance (b=0.13, p_{MCMC}<0.001) with P3 amplitude decreasing with distance from the relevant target endpoint. There was also a main effect of target (b=-0.37, p_{MCMC}=0.002) with larger P3s for *axe*, as well as a bound-
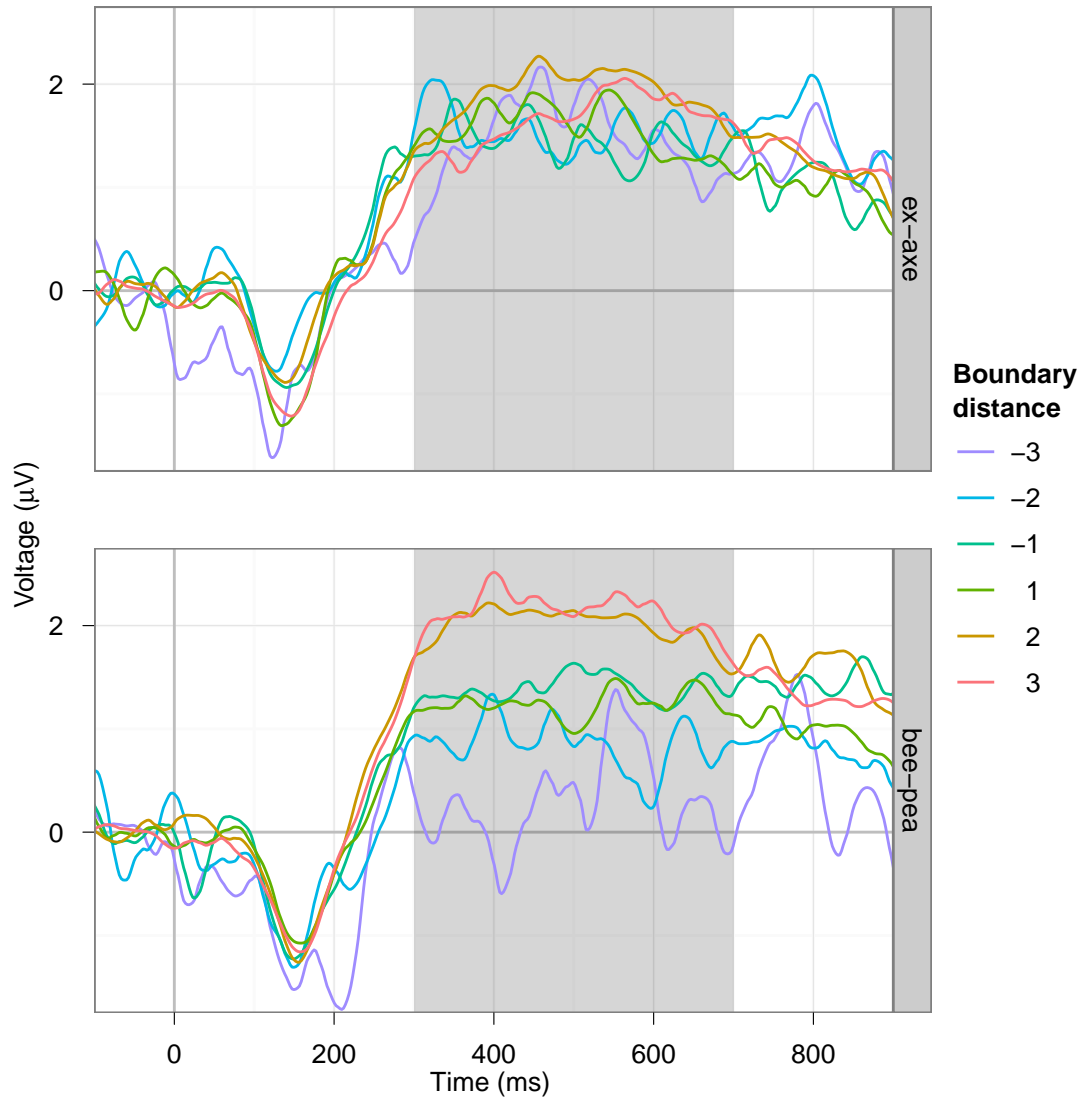
Figure 4.27: Experiment 6 results — ERP waveforms for parietal channels by category boundary distance on target-response trials. The top panel shows ERPs for the *ex-axe* stimuli, and the bottom panel shows ERPs for the *bee-pea* stimuli. Positive step numbers indicate steps on the target side of each participant's category boundary. Shaded areas indicate time range for the P3.
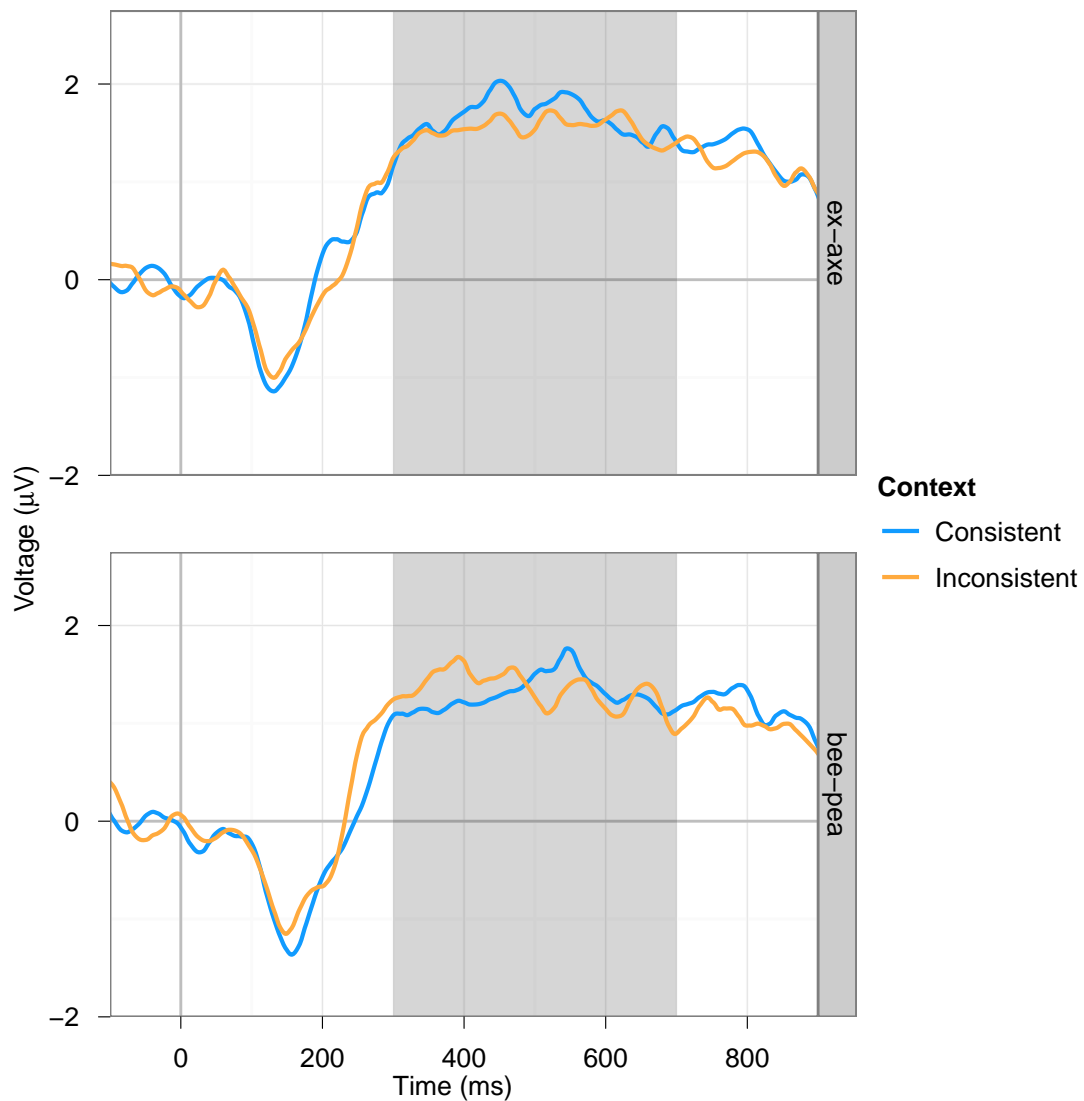
Figure 4.28: Experiment 6 results — ERP waveforms for parietal channels by context on target-response trials. The top panel shows ERPs for the *ex-axe* stimuli, and the bottom panel shows ERPs for the *bee-pea* stimuli. The data are averaged across the rStep conditions shown in Figure 4.27. Shaded areas indicate time range for the P3.
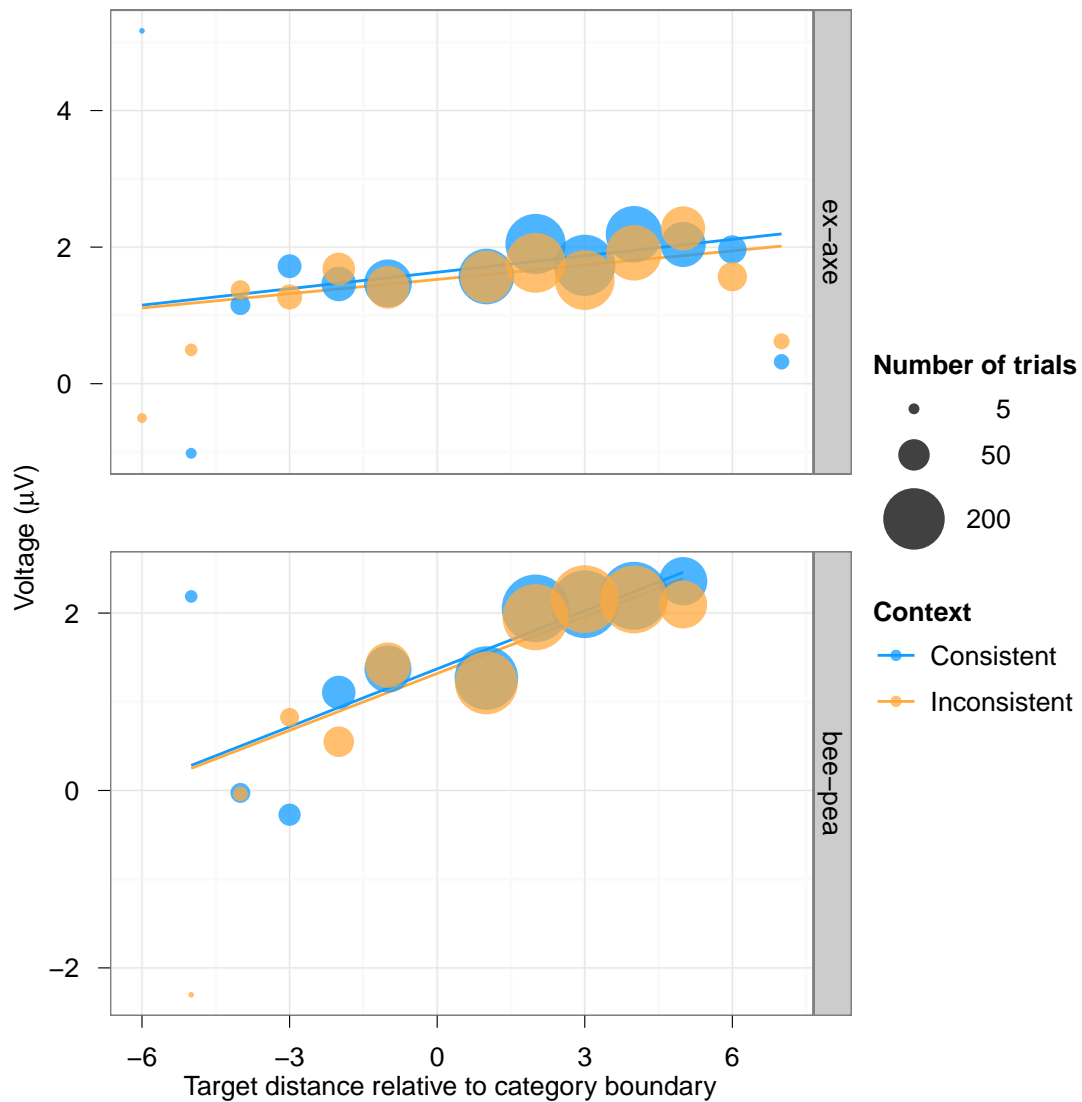
Figure 4.29: Experiment 6 results — P3 amplitude by category boundary distance and context on target-response trials. Mean amplitude as a function of stimulus continuum, distance from participants' category boundaries, and preceding context. The top panel shows ERPs for the *ex-axe* stimuli, and the bottom panel shows ERPs for the *bee-pea* stimuli. Larger category boundary distances indicate steps closer to the target endpoint. The size of the data points is proportional to the number of trials in each condition, and lines represent weighted linear models.

ary distance x target interaction (b=-0.13, p_{MCMC}=0.021). A follow-up analysis found a significant effect of boundary distance for both targets (*ex*: b=0.080, p_{MCMC}=0.045; *axe*: b=0.20, p_{MCMC}<0.001). The main effect of context and the other interactions in the original model were nonsignificant (context: b=0.0025, p_{MCMC}=0.999; distance x context: b=-0.050, p_{MCMC}=0.362; target x context: b=0.16, p_{MCMC}=0.498; distance x target x context: b=0.026, p_{MCMC}=0.821). Overall, these results are consistent with previous experiments showing sensitivity to within-category acoustic information.

4.5.3   Discussion

The results of this experiment provide tentative support the hypothesis that listeners encode cue-values relative to expectations. This was observed for gender differences as a function of VOT, but not for F1. Thus, this suggests that listeners' compensate for contextual variability in talker gender at the level of cue encoding: in the context of a female talker, VOT values were treated as shorter, compensating for the fact that women have longer VOTs than men. This provides a possible mechanism for the behavioral effects observed in listeners' categorization responses for the VOT stimuli in the context of the two talkers.

This conclusion is somewhat tentative, however, since a corresponding effect was not found for the F1 continuum. For these stimuli, the relative encoding hypothesis predicts that N1 amplitude should be larger in the context of a male talker than in the context of a female talker (since men have lower F1s than women and previous experiments showed that N1 amplitude increases with increasing F1). An effect on the N1 was not observed, even though listeners' compensated for the gender difference, as seen in their overt responses.

There are several reasons why an effect of talker gender may not have been seen in the N1. First, listeners may encode some cues relative to context (e.g., VOT) but not others (F1). However, this explanation seems unlikely given the considerable overlap between formant frequencies for different vowels (compared to the small amount of overlap in VOT for stops) and the fact that talker variability accounts for a large proportion of the variance

in F1 (Cole et al., 2010), both of which suggest a greater need for context compensation.

A second possibility is that, because the effects of F1 differences on N1 amplitude are weaker than those for VOT, we may not have had enough power to see a context effect. This is supported both by the weaker effects seen for F1 in Experiment 4 and the lack of any effect of F1 in this experiment. Given that we were unable to observe encoding of F1 in this experiment, we would not necessarily expect to observe context effects on F1 either.

Third, there could be an overall effect of talker gender that affected listeners' processing of both cues similarly. Specifically, if N1 amplitude is larger overall in the context of a female talker than a male talker, regardless of the stimulus continuum, the effect of context compensation in the F1 stimuli may have been negated, since it predicts larger N1s for the male talker. In contrast, larger N1s for the female talker would not cancel out the effects of context compensation for the VOT stimuli — both produce larger N1s for the female talker. Indeed, this fits with the pattern of results that was observed. An effect of talker gender was found for VOT, but no effect was found for F1 differences. Thus, the properties of the stimuli and direction of N1 differences may have masked effects of relative encoding of F1.

This experiment also provides additional evidence that listeners are extremely sensitive to small acoustic differences during category-level processing, as seen in the results of the P3 response for both stimulus continua. This replicates our earlier ERP results (Toscano et al., 2010) and extends behavioral and eye-tracking results, which have primarily been shown for voicing differences (McMurray et al., 2002), to vowel continua. In addition, there is some evidence indicating the preceding context affects P3 amplitude, though the effect was only observed for the F1 continuum (even though both continua had an effect on listeners' responses) and was not consistent across the entire range of F1 values. Thus, although differences in P3 amplitude may provide an additional measure of context compensation at later stages of processing, it may be difficult to observe these relatively small effects.

Overall, these results provide preliminary evidence in favor of relative cue encoding

based on extrinsic context information. In the section below, the predictions of the context compensation models are evaluated in terms of the results of this experiment and the other three experiments in this chapter.

4.6   General discussion

The results of these experiments allow us to distinguish between several models of context compensation. First, Experiments 5 and 6 argue against purely intrinsic accounts of talker compensation (Syrdal & Gopal, 1986; Christovich & Lublinskaya, 1979). Despite the fact that the relationships between formants (the source of information for intrinsic talker compensation for vowels) were constant in the target words, listeners still showed an effect of talker gender based on information in the preceding carrier phrase. Moreover, an effect on voicing for the *bee-pea* stimuli provides even stronger evidence against intrinsic approaches. Thus, these results rule out completely intrinsic accounts, thought they would still allow for models that combine extrinsic and intrinsic approaches (Nearey, 1989).

Second, Experiment 6 provides some evidence in favor of an extrinsic encoding process for handling talker variability. This is seen in the N1 differences to the *bee-pea* stimuli varying in VOT, where larger N1s are observed in the context of a female talker, and smaller N1s are observed in the context of a male talker, consistent with effect of VOT on N1 amplitude seen previously and the shifts in listeners categorization responses for these stimuli. However, a corresponding effect was not found for the F1 stimuli. This may have been due to the relatively small amount of variation in the N1 for these stimuli. Thus, the conclusion that listeners encode cues relative to preceding context is still tentative.

These experiments also replicate the results of Toscano et al. (2010) showing a linear effect of VOT on N1 amplitude that does not reflect categorical differences. In addition, they suggest that we can see effects of formant frequency differences on N1 amplitude. Experiments 1 and 4 show an effect of vowel differences on the N1, with /æ/ showing larger N1s than /ɛ/. Experiments 4 and 6 specifically varied these stimuli along an F1 continuum

to see if an effect of F1 encoding independent of listeners' phonological categories could be found. There was an overall linear effect for the F1 stimuli in isolation (Experiment 4), such that N1 amplitude is larger for higher frequency F1-values (i.e., consistent with an /æ/). Moreover, there was no evidence to indicate an effect of listeners' phonological categories. Thus, these results suggest that N1 amplitude can serve as an index of cue encoding for F1 differences as well as VOT differences, though the results are not particularly strong.

Overall, these results are most consistent with an extrinsic encoding process for handling talker variability. However, they do not provide information about the precise mechanism by which context information influences cue encoding. Models like C-CuRE (McMurray & Jongman, 2011) and the results of experiments using abstract, non-auditory talker information (Strand & Johnson, 1996) predict that this occurs via feedback from higher-level representations. In contrast, accounts like auditory contrast (Lotto & Kluender, 1998) suggest that lateral interactions between cue-level representations (in this case, continuous acoustic information about the talker from the carrier phrase) explains this effect. The next set of experiments examines this distinction by looking at speaking rate differences, which provide listeners with continuous acoustic information from context and could be used in a lateral model, but do not activate distinct categories, which would be necessary for some types of feedback models.

CHAPTER 5
SPEAKING RATE

5.1 Background

The next set of experiments examines the effect of speaking rate context on voicing judgments (e.g., /b/ vs. /p/). Along with talker variability, speaking rate has long been discussed as a major source of the lack of invariance in speech. Various models for handling speaking rate effects have been proposed, and while these draw on the same general classes of approaches used to handle talker variability, they are typically designed specifically for speaking rate. Again, these models can be distinguished based on the type of encoding they propose and how context information is combined with acoustic cues for voicing. However, unlike talker effects, variability in rate does not give listeners information about distinct categories (e.g., male vs. female). If feedback from categories is used for compensation, we would not expect an effect of continuous context information, like rate, on cue encoding. In contrast, models that encode relative cues using direct interactions between context information and phonetic cues (e.g., auditory contrast approaches) would predict effects of rate on cue encoding, in addition to effects of talker. Thus, some models make different predictions for rate context than for talker context.

The present set of experiments will test the predictions of these models, evaluating the proposals given by intrinsic, extrinsic, and raw-cue encoding approaches, and focusing more closely on the predictions given by lateral and feedback approaches. Before presenting the experiments, I will present an overview of phonetic and perceptual data on rate effects, as well a discussion of the compensation approaches that have been proposed to account for this data. Crucially, the available data does not address the effects of rate during cue encoding, motivating a need for work using the ERP paradigm presented here.

5.1.1   Phonetic data

Most languages distinguish between stop consonants using two or three voicing contrasts (Lisker & Abramson, 1964; Cho & Ladefoged, 1999). In English, voiced (or unaspirated) sounds like /b/ and /d/ are distinguished from voiceless (or aspirated) sounds like /p/ and /t/. A large number of acoustic cues that distinguish these sounds have been identified. Lisker (1986), for example, identifies 16 cues for word-medial voicing distinctions in English, including vowel duration, closure duration, F0, release burst intensity, and F1 transition duration, among other cues.

In syllable-initial position, voice onset time (VOT) is an extremely reliable cue that distinguishes voicing categories in most languages (Lisker & Abramson, 1964). VOT is defined as a time difference between two articulatory events (the release of the closure and the onset of glottal voicing). As a temporal cue, it is highly affected by factors like speaking rate, such that at fast speaking rates, VOT values tend to be closer to zero than at slow speaking rates (Allen & Miller, 1999; Kessinger & Blumstein, 1997; Beckman, Helgason, McMurray, & Ringen, 2011). In English and other languages that use aspiration, speaking rate primarily affects the voiceless category (/p/) which is marked by both longer and more variable VOTs than the voiced (/b/) category (Allen & Miller, 1999; Kessinger & Blumstein, 1997). For example, Allen and Miller (1999) found that when subjects were asked to speak quickly, they produced shorter VOTs than when asked to speak slowly. Other studies have also found differences in VOT as a function of speaking rate (Beckman et al., 2011; Kessinger & Blumstein, 1998; Miller, Green, & Reeves, 1986; Pind, 1995). These differences in VOT produce changes in the boundary between voiced and voiceless sounds, something listeners must compensate for to accurately recognize stop consonant voicing.

Speaking rate has effects on cues in both the same syllable as the consonant and other syllables, suggesting that both extrinsic and intrinsic rate compensation processes may be engaged by listeners. Extrinsic information comes primarily from the preceding

sentence rate (SR; Summerfield, 1981; Wayland, Miller, & Volaitis, 1994). The primary intrinsic cue is the length of the subsequent vowel, which covaries with speaking rate in the same sense and can be used for intrinsic encoding in the same way that that formant ratios are used for intrinsic compensation in vowel sounds. For example, the ratio between VOT and vowel length (VL) may offer a more invariant cue to voicing than raw VOT (Boucher, 2002). VL may also simply serve as a secondary cue for voicing, allowing feedforward accounts to explain these effects (McMurray, Clayards, et al., 2008; Toscano & McMurray, 2011a, 2010). There has also been debate about the relationship between VOT and VL in phonetic measurements (Kessinger & Blumstein, 1998; Allen & Miller, 1999), partly due to differences in the definition of VL (Turk, Nakai, & Sugahara, 2006). For the present set of experiments, I will focus on variation in preceding SR, rather than VL, since the effects of interest concern the initial encoding of VOT.

5.1.2 Perceptual data

A number of studies have demonstrated that rate, including both the effect of speaking rate in running speech and the effect of VL in individual words, has an effect on phonological judgments (Miller & Liberman, 1979; Summerfield, 1981; Miller & Dexter, 1988; Pind, 1995; Boucher, 2002; Miller & Volaitis, 1989; McMurray, Clayards, et al., 2008; Miller & Wayland, 1993; Toscano & McMurray, 2011a; Repp & Lin, 1991). Summerfield (1981) examined both effects of preceding sentence rate (SR) and length of the following vowel on listeners voicing judgments. As expected, listeners consistently identified stimuli with VOTs near 0 ms as voiced and VOTs near 40 ms as voiceless. Crucially, he also found a shift in listeners' VOT boundary as a function of differences in SR and VL. In both cases, faster rates produce more /p/ responses and slower rates produced more /b/ responses. However, the effect of VL was much larger than the effect of SR, leading Summerfield conclude that intrinsic compensation may be more important.

These results have been replicated a number of times using synthetic speech (Miller

& Dexter, 1988; Pind, 1995; Miller & Volaitis, 1989; McMurray, Clayards, et al., 2008; Miller & Wayland, 1993; Toscano & McMurray, 2011a; Repp & Lin, 1991). However, there has been debate about whether rate information also affects voicing judgments in natural speech. Shinn, Blumstein, and Jongman (1985) examined the effect of speaking rate on a manner of articulation (/b/-/w/) distinction that generally shows rate effects similar to VOT (Miller & Liberman, 1979). They modified the /b/-/w/ stimuli from Miller and Liberman (1979) by varying additional cues to manner in order to create more natural-sounding speech. They found that this eliminated the effect of rate on listeners' manner judgments and suggested that, under realistic listening conditions, listeners do not use rate information. Miller and Wayland (1993) demonstrated that when these natural-sounding stimuli are presented in the presence of background noise (presumably creating more natural conditions), an effect of rate is observed, suggesting that rate may be used in natural speech under more challenging listening conditions.

There has been less work examining effects of VOT in natural speech, but Utman (1998) examined listeners' responses to natural speech and found no evidence of rate effects (indicated by VL) on voicing judgments. However, recent studies using more sensitive paradigms have been able to find small effects of VL on voicing in natural speech (Boucher, 2002; Toscano & McMurray, 2011a) Thus, although the effect of rate may be smaller in natural speech, listeners are sensitive to VL and can use it to make phonological judgments.[1] However, previous studies have not demonstrated an effect of preceding SR in natural speech. Given that this distinguishes extrinsic and intrinsic compensation accounts, and the debate about whether there are any rate effects at all in natural speech, an important goal of this set of experiments is to establish that there are effects of preceding SR in natural speech and then to examine whether such effects are driven by cue-level or category-level compensation.

_____

[1]See Toscano and McMurray (2010) for a discussion of why rate effects would be smaller in natural speech than in synthetic speech.

5.1.3   Mechanisms for handling rate variability

Similar to formant ratios used for talker variability compensation, several researchers have suggested that the ratio between VOT and VL is invariant across rate contexts (Port & Dalby, 1982; Kessinger & Blumstein, 1998; Boucher, 2002), and can serve as a source of information for intrinsic rate encoding. Some phonetic data supports this. Kessinger and Blumstein (1998) found that the VOT:VL ratio remained constant for voiceless stops across different speaking rates for a given vowel context. Similarly, Boucher (2002) found that the ratio boundary between voiced and voiceless stops was constant across speaking rates.

A number of perceptual studied have also suggested listeners compute context-independent ratios between cues. Port and Dalby (1982) replicated the early results of Denes (1955) and suggested that the ratio between the vocalic and frication portions of a vowel-fricative syllable serves as an invariant cue to word-final fricative voicing. Similarly, Boucher (2002) found that boundaries computed from VOT:VL ratios in his production data predicted listeners responses in a categorization task in which the ratio was varied by holding VOT constant at an ambiguous value and changing VL.

However, results of other studies challenge VOT:VL ratios as a mechanism for handling variability in VOT caused by rate. Pind (1995) collected production and categorization data from Icelandic speakers and found that the ratio boundary predicted from production data did not match listeners' perceptual boundaries, challenging this approach and the cross-linguistic generality of earlier results.

Other work has argued against an intrinsic VOT:VL ratio model by demonstrating that listeners show independent effects of VOT and VL. Crucially because word-initial cues, like VOT and formant transition durations, and VL are separated in time, intrinsic models suggest that listeners would have to wait until both VOT and VL arrive before they can make a voicing judgment. Therefore, these models predict that the effect of context information, like VL, and a primary phonetic cue, like VOT, should occur at the same point in time, since

the relevant cue for phonological distinctions is the ratio between the primary phonetic cue and VL. Thus, phoneme and lexical categories cannot be activated until listeners have information from both that cue and VL. In contrast, raw-cue encoding models predict that VL serves as a weak phonetic cue. In this case, the effect of VOT would precede the effect of VL (since VOT precedes VL information in time).[2]

Miller and Dexter (1988) first looked at this by examining listeners' earliest responses in a speeded-response task. They found that listeners' could make word-initial voicing judgments on the basis of VOT without information about VL. Recently, several experiments using the visual world eye-tracking paradigm (VWP; Tanenhaus, Knowlton, Eberhard, & Sedivy, 1995; McMurray et al., 2002) have extended this using a more online measure (McMurray, Clayards, et al., 2008; Toscano & McMurray, 2011a). In VWP tasks, listeners hear spoken words or sentences corresponding to pictures on a computer screen or real objects, while their eye-movements are recorded. Listeners will make eye-movements to an object before moving the cursor to it (in the case of a computer display) or manipulating it with their hand (in the case of real objects). The proportion of eye-movements to each item provides a measure of how strongly the participant is considering that item (Allopenna, Magnuson, & Tanenhaus, 1998; Dahan, Magnuson, & Tanenhaus, 2001; Dahan, Magnuson, Tanenhaus, & Hogan, 2001; Dahan & Gareth Gaskell, 2007; Magnuson, Dixon, Tanenhaus, & Aslin, 2007) and has been used to show effects of both VOT (McMurray et al., 2002) and VL (Salverda, Dahan, & McQueen, 2003). Importantly, listeners' eye-movements at a given point in time only reflect processing up until that point, allowing us to directly assess whether VOT and VL are being processed together (i.e., at the same point in time) or independently (at different points).

McMurray, Clayards, et al. (2008) tested this by presenting listeners with /b/-/p/ and

---

[2]Note that this only applies to models in which compensation is required at the time of initial cue encoding. Models that allow for initial encoding of raw cues could still show independent effects of VOT and VL.

/b/-/w/ minimal pair words varying in VOT (for /b/-/p/) or formant transition duration (for /b/-/w/) and VL. They found that the effect of VOT or FTD occurred approximately 100 ms before the effect of VL. Thus, listeners can use VOT independently of VL. This argues against intrinsic compensation models and suggests listeners do not use the ratio between VOT and VL to make voicing judgments. Toscano and McMurray (2011a) replicated this result using natural speech.

Based on these results, Toscano and McMurray (2010) suggested that rate compensation could be accomplished in a purely feedforward model if VL is simply treated as a secondary cue to voicing. They simulated speaking rate effects using a weighted Gaussian mixture model (WGMM; a feedforward, raw-cue encoding model). The model was trained on distributions of VOT and VL values for voiced and voiceless sounds in English and showed that, as with human listeners, the model's responses reflected effects of both VOT and VL on voicing judgments. In addition, this was accomplished without encoding VOT relative to VL. Rather, in this model, both cues simply bias voicing categorization (albeit with less bias due to weaker information provided by the distributional statistics of VL). Simulations with the model also replicated differences in the size of the VL effect seen between natural and synthetic speech for human listeners, suggesting that even this apparent reweighting of cues can be accomplished in a purely feedforward system without changes in cue weights. Thus, it is possible to treat VL as a phonetic cue rather than as a context effect, and raw-cue encoding via feedforward activation may be able to account for at least some rate effects.

These results provide information about how listeners compensate for differences in VL, but they do not address the contributions of preceding SR. Listeners did not have information about preceding SR in these experiments, so the results could not be attributable to extrinsic cue encoding. Given this, and the independent effects for VOT and VL, the results so far are most consistent with a raw-cue encoding approach, though it is unclear

whether a similar approach would work for SR — it seems unlikely that specific SRs are directly associated with voiced and voiceless categories. Moreover, it is unclear whether an extrinsic rate compensation approach could take advantage of feedback from category-level information (which can be used for talker compensation). Because rate differences are continuous, listeners may not have distinct categories against which they can compute relative phonetic cues. In contrast, talker gender does give listeners categories that can be used to adjust cues. In addition, the results of work on talker compensation suggest that relatively abstract information about talker gender (such as a picture of a face; Johnson et al., 1999) can serve as a source of information for compensation via feedback to cue-level encoding. Thus, if listeners encode cues relative to context using a feedback mechanism, we may not expect to see an effect of SR on cue-level representations. Rather, if listeners encode cues relative to SR, it would suggest they use a lateral, extrinsic compensation approach.

5.1.4 Experiment overview and predictions

The various compensation approaches make different predictions about whether and how SR affects cue encoding. Purely intrinsic compensation and raw-cue encoding accounts, predict that SR has no effect on VOT encoding, and thus no effect on the N1. Intrinsic compensation approaches suggest that VOT is processed in relation to VL rather than SR (thus, SR should have no effect). Results like those of Summerfield (1981) rule out such models for synthetic speech, but, given previously observed differences in rate effects between natural and synthetic speech, it is not clear if these results extend to natural speech. Raw-cue encoding approaches allow for the possibility of SR effects at later stages of processing, but also predict no effect on cue encoding. In both cases, these two approaches predict that VOT will be encoded veridically when VL is held constant but preceding SR varies. Thus, they make the same predictions that they made for the talker identity experiments.

Extrinsic compensation accounts make different predictions about whether VOT is encoded relative to preceding rate. Models that use feedback from category-level repre-

sentations to compute relative cue-values do not predict effects of rate on VOT encoding, since variation in rate does not activate specific categories (unlike variation in talker gender). Thus, we can likely rule out such models as plausible accounts of compensation for rate. In contrast, models that compute relative cues solely from information at the level of cue encoding (lateral models) do predict an effect of preceding SR on VOT encoding. Thus, if we see an effect of context compensation on the N1, it is most likely due to lateral interactions. As a result, while the talker identity experiments allowed us to distinguish between extrinsic encoding and other approaches, this set of experiments will allow us to distinguish between these two extrinsic encoding approaches.

Before directly assessing the question of whether SR affects VOT encoding, we must first determine that the SR in naturally-produced stimuli has an effect on listeners' voicing judgments as there has been debate in the literature about whether listeners compensate for rate differences in natural speech (Shinn et al., 1985; Miller & Wayland, 1993; Utman, 1998). Experiment 7 will address this point by presenting listeners with sentences varying in SR, VOT, and VL. If SR does have an effect, the phonetic data predict that listeners will show more voiceless responses in the context of fast speech. This result would rule out a purely intrinsic compensation account (since an extrinsic factor [SR] has an effect), but would be consistent with models that combine extrinsic and intrinsic compensation (Nearey, 1989), as well as raw-cue encoding models (e.g., feedforward models that combine SR with phonetic cues at later stages of processing).

Next, Experiment 8 tests the predictions made by different types of context compensation accounts by presenting listeners with stimuli varying in VOT and preceding SR, and measuring N1 and P3 responses using the same approach as Experiment 6. If SR has an effect on VOT encoding (i.e., N1 responses), it would support a lateral, extrinsic encoding model. If SR does not have an effect on VOT encoding but does have an effect on listeners' responses, it would support raw-cue encoding approaches, though it would not rule out

an additional extrinsic compensation process that uses feedback from categories to handle other types of context effects (since feedback from categories simply could not be used in this situation). Finally, if SR does not have an effect on VOT encoding or overt responses, it would support a purely intrinsic compensation approach or a raw-cue approach that does not use SR.

5.2 Experiment 7: Effect of rate on voicing judgments

In this experiment, the effect of SR on listeners' voicing judgments in natural speech was examined. Several studies have suggested that listeners do not use rate information for voicing or manner judgments in natural speech (Shinn et al., 1985; Miller & Wayland, 1993; Utman, 1998). However, it may be that these effects are simply smaller than in synthetic speech. Toscano and McMurray (2010, 2011a) demonstrated that this was the case for rate information cued by VL, but this has not been examined for SR. Thus, before examining context effects in an ERP experiment, we must first establish that variation in SR influences voicing in natural speech.

Data from this experiment were collected as part of a larger project examining listeners' use of rate information during online spoken word recognition using an eye-movement measure of lexical activation (Toscano & McMurray, 2011b). Both eye-movement measures and listeners' categorization responses were collected, but only the categorization responses are reported here since they are sufficient for determining whether SR has an effect on voicing judgments in natural speech. We are primarily interested in the effects of SR here, but VL was included to determine if other effects of rate could be observed.

5.2.1 Methods

5.2.1.1 *Participants*

Twenty people participated in the experiment. Participants were recruited from the University of Iowa community according to University human subjects protocols, provided

informed consent, and were either compensated $40 or received course credit for partici-
pation. Participants reported English as their only native language, normal hearing, and
normal or corrected-to-normal vision.

5.2.1.2 Design

Participants performed a 4AFC picture identification task. Auditory stimuli con-
sisted of the six sets of /b/-/p/ minimal pair target words used in Experiment 2 (*bath-path*,
beach-peach, *beak-peak*, *bet-pet*, *bike-pike*, and *buck-puck*) preceded by one of five carrier
sentences (Table 5.1). Target words varied along nine-step VOT continua and across two
vowel length (VL) conditions, with the carrier sentence spoken at either a fast or slow rate.
Each minimal pair was grouped with two unrelated words starting with /l/ and /ʃ/. For each
participant, an /l/ and /ʃ/ item was randomly assigned to a particular minimal pair with
the requirement that semantically related words (e.g., *pet* and *leash*) could not be paired
together. /l/ and /ʃ/ items also varied in VL and were preceded by the carrier sentences
varying in SR. Stimuli were presented in random order, each target stimulus was repeated
three times, and there were 540 unrelated trials for a total of 1080 trials. The experiment
was conducted over the course of two days and lasted approximately 90 minutes each day.[3]

5.2.1.3 Stimuli

Stimuli were created from recordings used in Experiment 2, along with recordings
of carrier phrases made during the same session. Carriers were recorded with a neutral
word at the end ("tongue") to avoid coarticulation.

VOT continua were created using the procedures described in Experiment 2 (see
Section 3.5.1.3). VL and SR conditions were created using the pitch-synchronous overlap-
add method to lengthen and shorten the sounds. For the two VL conditions, the onset and

---

[3]Due to an error in the experiment presentation code, participants only ever heard either the /ʃ/
or /l/ item in a given item-set. Although this affected their eye-movements to objects in the display,
it does not appear to have had any effect on their mouse-click responses to the /b/ and /p/ words.

Table 5.1: Carrier phrases used in Experiment 7.

Carrier Phrase
1. On this screen, click on the...
2. In this display, choose the...
3. On this screen, select the...
4. In this display, pick the...
5. On this screen, please choose the...

offset of the vowel was marked for each sound, and VLs were increased or decreased by 40% of their original duration. Carrier sentences were increased and decreased by 15% of their original duration to create the slow and fast SR conditions, respectively. These differences in sentence length produced speaking rates in a similar range to those reported by Miller, Grosjean, and Lomanto (1984). Finally, each carrier sentence was spliced onto each target and unrelated stimulus.

Visual stimuli were clipart images that were selected and processed using standard lab procedures that have been used for several prior studies (e.g., McMurray, Samelson, Lee, & Tomblin, 2010). For each word, several pictures were downloaded from a clipart database. A team of researchers selected the picture that depicted the most canonical representation of the word and edited it if needed (e.g., to remove objects in the background).

5.2.1.4 Procedure

Participants were seated in the same room used in Experiment 5 and wore an SR Research Eyelink II head-mounted eye-tracker. The eye-tracker was calibrated using the standard 9-point calibration grid, and, after calibration, participants began the experiment. First, two sets of training trials were presented to familiarize participants with the pictures and words. In the first part of training, participants saw each picture in the center of the

screen. The written word corresponding to the picture appeared below the image 500 ms later. After viewing the picture and reading the name, participants clicked the mouse button to continue to the next trial. Each picture was presented once in random order.

In the second part of training, four pictures (one set of /b/, /p/, /l/, and /ʃ/ items) appeared in the four corners of the display, and the written word corresponding to one of the pictures appeared in the center of the display 500 ms later. Participants clicked on the picture corresponding to the written word to go onto the next trial (only clicking the correct picture allowed them to continue). Each picture set was presented twice in random order, and the arrangement of the pictures in the display was randomized.

After training, the experimental trials began. Each trial proceeded similarly to the second part of training. One set of four pictures appeared in each corner of the display, and a blue circle appeared in the center. After 500 ms, the circle turned red, participants clicked on it, an auditory stimulus was played over the headphones, and participants made their response by clicking on the picture corresponding to the instruction they heard. The arrangement of the pictures was randomized such that, within an experimental condition, each relative arrangement of minimal pair items (i.e., adjacent horizontally, adjacent verti-cally, or diagonal) occurred equally often.

5.2.2 Results

Listeners correctly identified the stimuli at the endpoints of the VOT continua (mean accuracy: 99.2%) and showed standard categorization functions. Figure 5.2 shows the pro-portion of /p/ responses as a function of VOT and SR. There were more voiced responses in the slow SR condition than in the fast SR condition, consistent with the prediction that preceding rate influences voicing judgments.

Responses were analyzed using a logit mixed-effects model with subject and word pair entered as random effects and VOT, VL, and SR as fixed effects (r_{max}=-0.224). The re-sult showed a main effect of VOT(b=1.80, z=50.32, p<0.001) with more voiceless responses

for long VOTs. There were also main effects of VL (b=-0.85, z=-11.45, p<0.001) and SR (b=-0.35, z=-4.77, p<0.001), such that listeners made more voiceless responses in the context of short vowels and fast sentences than in the context of long vowels and slow sentences. Thus, SR has an effect on voicing in natural speech. These results fit with the predictions from the phonetic data, showing that listeners produce more voiceless responses in in the context of fast speech (compensating for the overall shorter VOTs). Finally, there was a VOT x SR interaction (b=-0.12, z=-2.02, p=0.043), indicating that the slope of the categorization function differed between the two SR conditions. None of the other interactions were significant (VOT x VL: b=-0.068, -1.19, p=0.232; VL x SR: b=0.01, z=0.07, p=0.945; VOT x VL x SR: b=0.15, z=1.36, p=0.175).

5.2.3   Discussion

These results show that SR has an effect on voicing in natural speech. This fits with the results of Toscano and McMurray (2011a) which showed that, in contrast to previous studies, effects of rate or weak secondary phonetic cues for voicing judgments can be seen in natural speech, though these effects may be smaller than in synthetic speech. Critically, for the present series of experiments, these results demonstrate that natural stimuli varying in SR can be used to examine speaking rate effects on VOT encoding and categorization.

Moreover, these results argue against purely intrinsic approaches that suggest listeners compensate for rate using within-word duration contrasts (like VOT and VL). Indeed, since both extrinsic (SR) and intrinsic (VL) information was available in this experiment, listeners could have relied solely on VL information to compensate for differences in rate. Instead, we found effects of both SR and VL. While this rules out completely intrinsic models, it is consistent with accounts that incorporate both extrinsic and intrinsic information (Nearey, 1989). Thus, along with the results of McMurray, Clayards, et al. (2008) and Toscano and McMurray (2011a) which show that VOT and VL are used independently, these results support extrinsic and raw-cue encoding approaches to handling rate effects.
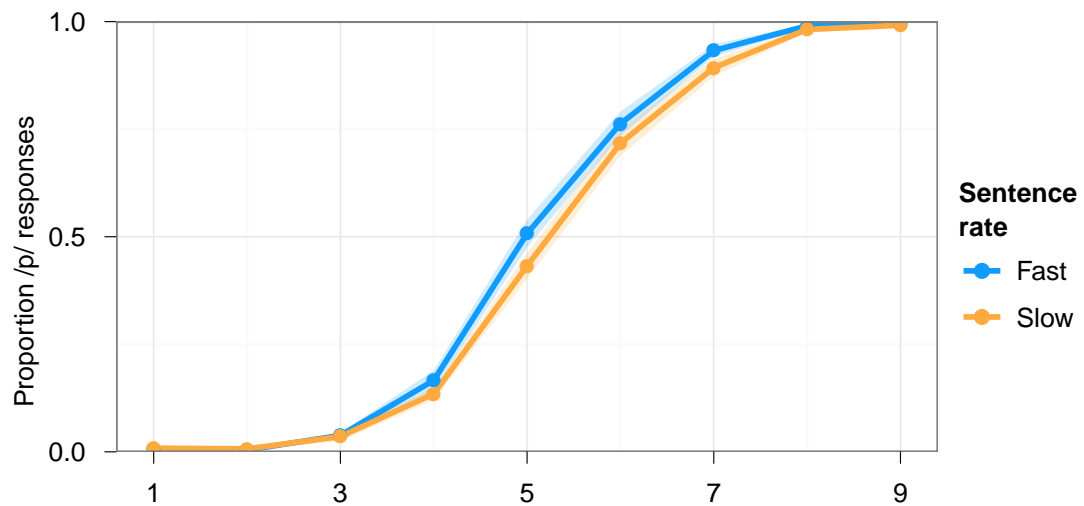
Figure 5.1: Experiment 7 results — categorization responses by sentence rate. Participants' mouse-click responses as a function of VOT for each SR condition. The fast SR condition produced more voiceless responses than the slow SR condition, indicating an effect of rate compensation on listeners' voicing judgments.
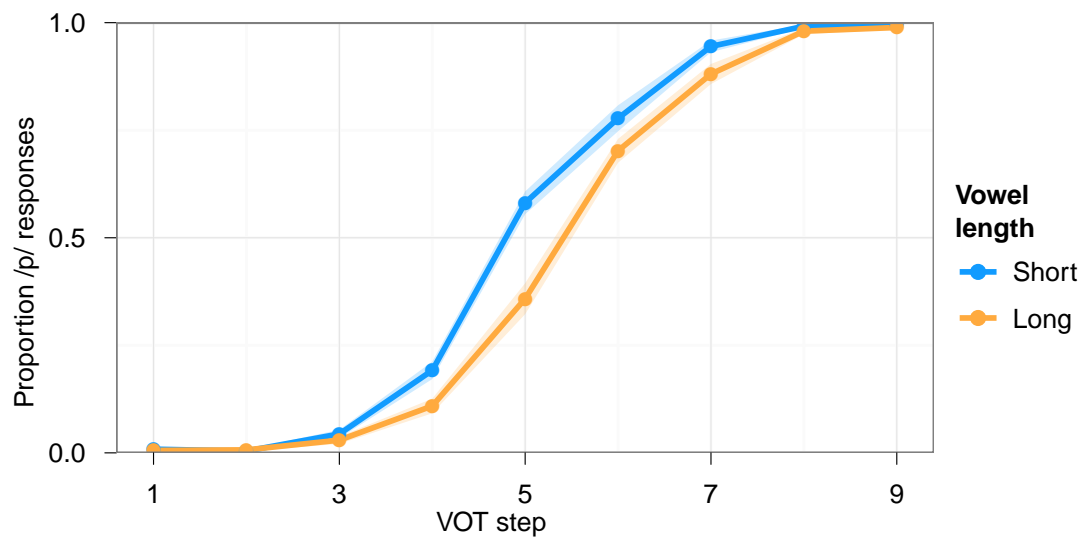
Figure 5.2: Experiment 7 results — categorization responses by vowel length. Participants' mouse-click responses as a function of VOT for each VL condition. The short VL condition produced more voiceless responses, consistent with compensation for rate differences.

5.3   Experiment 8: Effect of rate on cue encoding

Together, the previous experiment and Experiment 2, which examined N1 responses to naturally-produced VOT differences, demonstrate that (1) preceding SR affects voicing judgments in natural speech, (2) differences in N1 amplitude can be observed for voiced and voiceless endpoints, and (3) a single N1 peak is observed for both. This experiment now tests the predictions of the two compensation mechanisms (feedback from categories vs. lateral interactions) by examining whether VOT is encoded relative to preceding SR. To do this, we measured ERP responses to words varying along a VOT continuum that were preceded by carrier phrases spoken at either a fast or slow rate.

Stimuli consisted of two continua: (1) a /b/-/p/ VOT continuum varying from *beach* to *peach* (the pair of words that produced the largest differences in N1 amplitude in Experiment 2), and (2) an /i/-/u/ vowel continuum varying from *ease* to *ooze*. This vowel distinction was chosen because it should produce differences in N1 amplitude along the continuum (based on pilot data) but should not be affected by rate, since vowel cues that are affected by rate (e.g., VL) do not reliably distinguish these sounds. The mean durations of /i/ and /u/ in English are 282 ms and 273 ms, respectively (Hillenbrand et al., 1995).

If preceding rate context affects VOT encoding, we would expect to see differences in N1 amplitude to the /b/-/p/ stimuli as a function of SR. Specifically, we would expect that VOTs in the slow SR condition (where the sound would be more likely to be perceived as voiced) would show a larger N1 (i.e., a response closer to those of the voiced end of the continuum). In contrast, if context does not affect cue encoding, we would not expect to see any differences in N1 amplitude for the SR conditions, but we may see an effect at later stages of processing (i.e., as variation in P3 amplitude). Specifically, we would expect that, for voiced targets, /b/-/p/ stimuli in the slow SR condition would show a larger P3, and, for voiceless targets, they would show a smaller P3. For the fast SR condition, P3 amplitude would be larger for voiced targets and smaller for voiceless targets. In addition, we do not

expect differences in N1 or P3 amplitude as a function of preceding SR for the /i/-/u/ stimuli, since rate does not affect the primary cues used to distinguish those vowels.

5.3.1 Methods

5.3.1.1 *Participants*

Twenty people participated in the experiment. Participant recruitment, consent, and compensation procedures were the same as Experiment 4, and participants met the same language, hearing, and vision criteria as in previous experiments.

5.3.1.2 *Design*

Participants performed a target detection task in which they identified whether stimuli matched a target word. Words were presented in the context of a carrier phrase that had either a slow or fast rate. In order to reduce overlap between ERPs to the carrier phrase and target word and to allow us to use the Adjar technique, the ISI between the onset of the carrier phrase and target word was varied over a 234 ms range in 10 steps by manipulating the duration of the carrier phrase within each SR condition. All of the ISIs in the fast SR condition were shorter than the ISIs in the slow SR condition. Each combination of continuum, VOT, SR, and ISI within the SR conditions was presented once in random order. Each of the four words served as the target in a different block, for a total of 864 trials.

As in Experiments 4 and 6, participants performed a 2AFC categorization task after the ERP session in order to get estimates of their category boundaries for each stimulus continuum. The procedure was the same as the one used in previous experiments. The experiment took approximately two hours and was completed in a single session.

5.3.1.3 *Stimuli*

New recordings were made for each continuum as well as the carrier sentences. The /b/-/p/ VOT continuum was created in the same way as Experiment 1. VOT values are listed

in Table 5.2. The mean difference between the actual cross-spliced VOT and the expected VOT for a given step along the continuum was less than 1 ms.

The /i/-/u/ continuum was created by changing the second and third formant frequencies (F2 and F3) and their bandwidths (B2 and B3) of the /i/ token to those of the /u/ token. Formants were measured and adjusted using linear predictive coding (LPC, Burg method; Andersen, 1974) with a time step of 10 ms, maximum formant frequency of 22.05 kHz, window length of 80 ms, 20 poles, and a pre-emphasis from 214 Hz. Table 5.3 lists the formant frequency and bandwidth values for the endpoints. Each step in the continuum used the same frication at the end of the word. This was created by sample averaging the frication from each token in order to minimize potential coarticulatory information. The final intensities of the frication and vocalic portions were set to 65 and 75 dB SIL, respectively, to produce a natural-sounding syllable. Finally, the words were low-pass filtered to 4410 Hz using a symmetric Hann filter with a smoothing width of 2000 Hz to remove artifacts caused by LPC resynthesis.

The carrier sentences were changed from those used in Experiment 7 so that the instruction was consistent with the task in this experiment ("On this trial, the word is..."). SR of the carrier sentences was manipulated in Praat using the pitch-synchronous overlap-add method (Moulines & Charpentier, 1990). Within each SR condition, the sentence length varied over a 234 ms range to create 10 different ISIs relative to the onset of the target word. Fast SRs varied from 1419 to 1653 ms, and slow SRs varied from 1961 to 2194 ms.

Table 5.2: VOT values for Experiment 8 /b/-/p/ continuum steps.

Continuum	1	2	3	4	5	6	7	8	9
beech-peach	0	6	12	16	20	25	29	36	40

Note: Values are in ms.

5.3.1.4 *Procedure, EEG recording, and data processing*

The experimental setup and data processing procedures were the same as in Experiment 4. The task was also the same but used the target words from the current experiment.

5.3.2 Results

5.3.2.1 *Behavioral results*

The first set of analyses examined listeners' overt responses to determine if an effect of SR could be seen for the /b/-/p/ continuum. Listeners' behavioral responses during the ERP session were converted from target/nontarget responses to responses corresponding to particular words (as in Experiments 4 and 6). For both continua, participants showed typical categorization functions and were highly accurate at the continuum endpoints (mean accuracy: 98.1%). However, there did not appear to be any differences due to SR for the VOT stimuli. Figure 5.3 shows listeners' responses for each stimulus continuum and SR.

Logit mixed-effects models were run for each continuum with step (VOT step for the *beach-peach* stimuli, formant frequency step for the *ease-ooze* stimuli), SR, and target (the two targets relevant to that continuum) as fixed effects (r_{max}=0.143). For the *beach-peach* stimuli, there was a main effect of VOT step (b=2.42, z=25.22, p<0.001). The effect of SR was not significant (b=0.15, z=0.81, p=0.417), suggesting that SR did not have an effect on their voicing judgments. This contrasts with the results of Experiment 7, though there are several differences between the two experiments that may have caused this (see discussion below).

Table 5.3: Formant values for Experiment 8 /i/-/u/ continuum endpoints.

Word	F2	F3	B2	B3
ease	2494	3473	86	497
ooze	899	2395	236	393

Note: Values are in Hz.
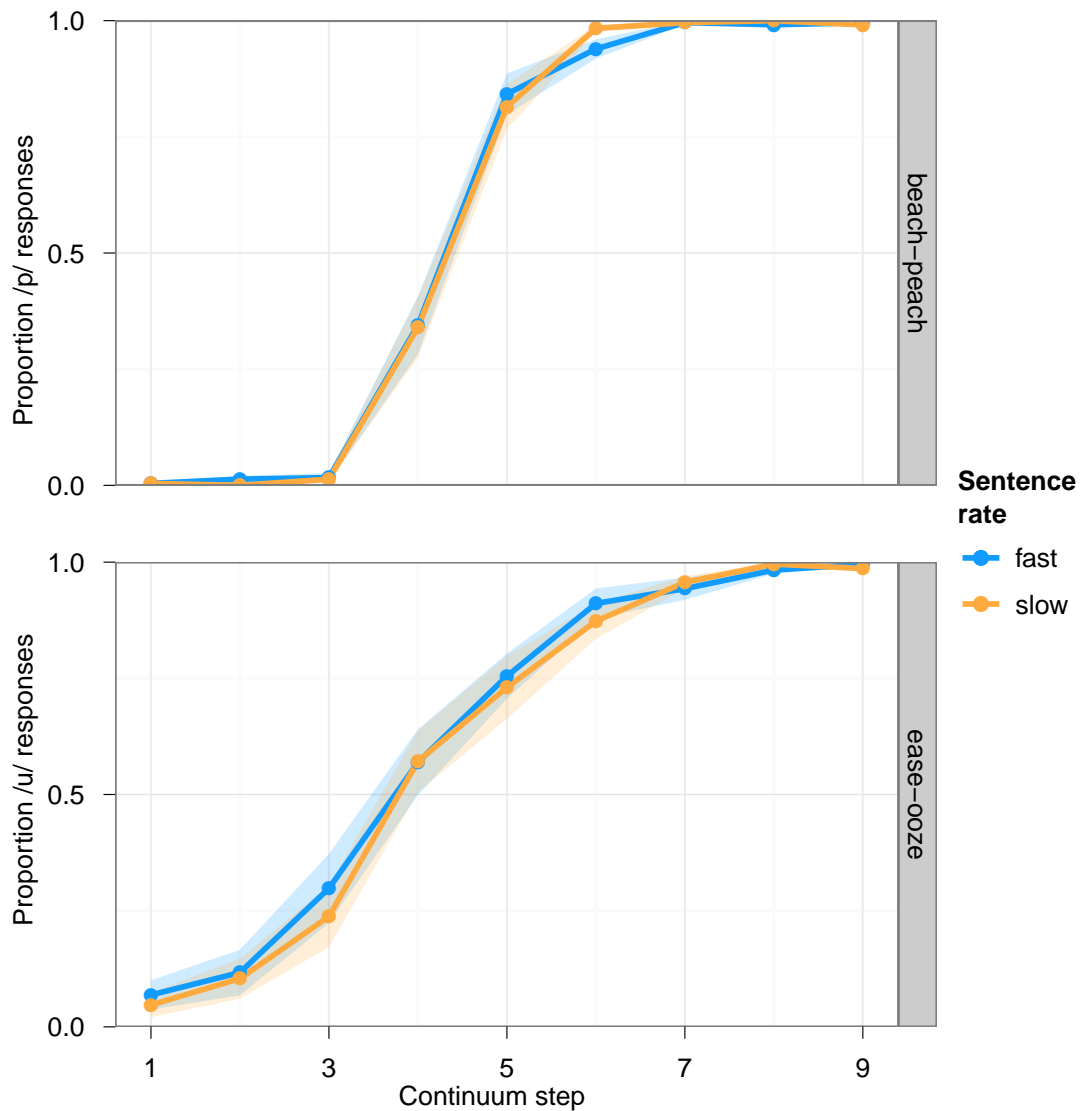
Figure 5.3: Experiment 8 results — behavioral responses during target detection task. Listeners' categorization responses during the target detection task as a function of continuum step and speaking rate for each stimulus continuum (top panel: *beach-peach* stimuli; bottom panel: *ease-ooze* stimuli).

There was a main effect of target (b=0.58, z=3.16, p=0.002), such that listeners made more "peach" responses when *peach* was the target and more "beach" responses when *beach* was the target, consistent with the previous target detection experiments. There was a marginal VOT x SR interaction (b=0.31, z=1.78, p=0.076) and a marginal VOT x target x SR interaction (b=-0.59, z=-1.68, p=0.094); all other interactions were nonsignificant (VOT x target: b=0.11, z=0.66, p=0.512; target x SR: b=-0.23, p=0.532).

A corresponding model was run for the *ease-ooze* stimuli (r_{max}=-0.136. There was a main effect of formant frequency (b=1.33, z=31.04, p<0.001), but no effect of SR (b=-0.15, z=-1.14, p=0.255) or formant frequency x SR interaction (b=0.062, z=0.88, p=0.377). This isn't surprising, since SR does not provide useful information for distinguishing these vowels. There was a main effect of target (b=0.59, z=4.52, p<0.001), again showing that listeners made more target responses than nontarget responses. None of the other interactions were significant (formant frequency x target: b=-0.032, z=-0.45, p=0.653; target x SR: b=-0.077, z=-0.30, p=0.766; formant frequency x target x SR: b=-0.23, z=-1.63, p=0.103).

As in Experiments 4 and 6, participants' category boundaries for each continuum were estimated by fitting their responses from the 2AFC categorization task that was run after the ERP session to four-parameter logistic functions. The mean category boundary for the *beach-peach* continuum was at step 4.8 (≈19 ms VOT), and the mean boundary for the *ease-ooze* continuum was at step 4.7.

5.3.2.2 *N1 amplitude*

As in Experiment 6, the Adjar procedure was used to reduce overlap between ERP components to the carrier phrase and those to the target words. As before, examination of waveforms for each ISI suggested that offset of the carrier phrase generated ERPs that occurred during the pre-stimulus baseline period for the target word. For the fast SR, ISI between carrier offset and word onset ranged from 130 to 142 ms. For the slow SR, ISI ranged from 176 to 196 ms. The Adjar procedure was applied and converged on a stable

estimate of the overlap for each SR condition (Figure 5.4).

Figure 5.5 shows grandaverage ERP waveforms as a function of continuum step, and Figure 5.6 shows waveforms as a function of preceding SR. As in previous experiments, N1 amplitude decreased with increasing VOT. In addition, there appeared to be an effect of SR on the N1 with larger N1s for fast SRs and smaller N1s for slow SRs.

N1 latency was later than for words in isolation, a result also seen in Experiment 6. Mean N1 amplitude was calculated for the average of the three frontal channels between 125 and 165 ms post-stimulus for the *beach-peach* stimuli and from 170 to 210 ms for the *ease-ooze* stimuli. Figure 5.7 shows N1 amplitude as a function of VOT and SR for each stimulus continuum.

To assess the effects of VOT and SR on the N1, mean amplitudes for the *beach-peach* stimuli were examined with a linear mixed-effects model with VOT step, SR, and target type as fixed effects (r_{max}=0). The model showed a significant effect of VOT (b=0.073, p_{MCMC}<0.001) with N1 amplitude decreasing with increasing VOT, replicating the same pattern observed in previous experiments.

There was a marginal effect of SR (b=-0.14, p_{MCMC}=0.064) with larger N1s for the fast SRs than for the slow SRs. This suggests there may be an effect of rate on VOT encoding, but it is not in the direction predicted by relative encoding models. Since N1 amplitude generally decreases with increasing VOT, and listeners would compensate for fast speech by computing VOT values as longer, we would expect a larger N1 in the context of a slow SR and a smaller N1 in the context of a fast SR. Thus, this effect may be due to an overall effect of SR on speech processing that is not specific to the acoustic cues being encoded.

The effect of target type was signifiant (b=-0.22, p_{MCMC}=0.005) with larger N1s when *ease* or *ooze* was the target. This is similar to the effect of target type observed in the previous experiments in which the nontarget trials produced larger N1s. None of the interactions were significant (VOT x target type: b=-0.033, p_{MCMC}=0.284; VOT x SR:
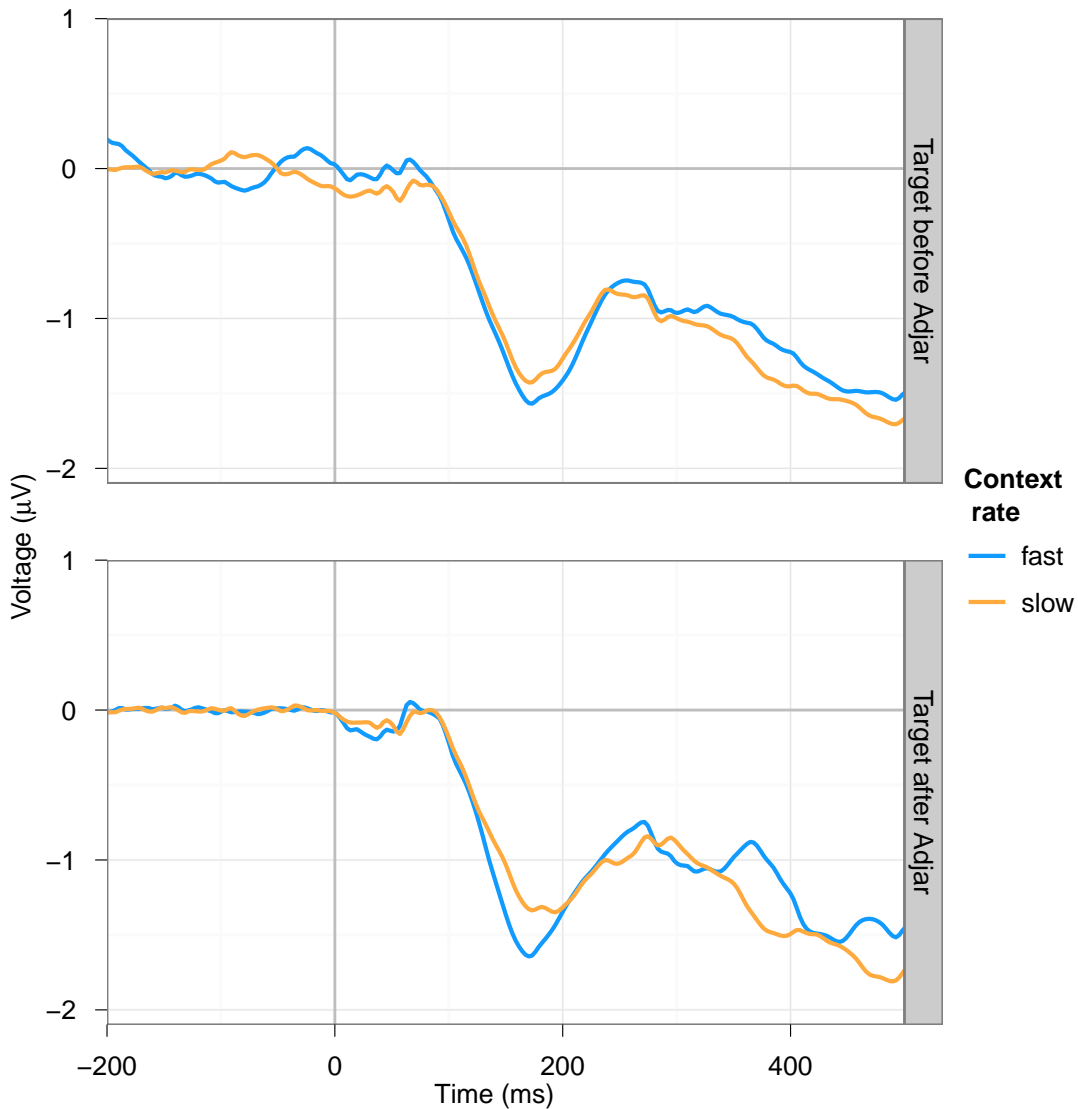
Figure 5.4: Experiment 8 results — effectiveness of Adjar procedure. Grandaverage ERP waveforms for the average of the three frontal channels before (top panel) and after (bottom panel) the Adjar procedure was applied. The amount of overlap was reduced for both speaking rate conditions.
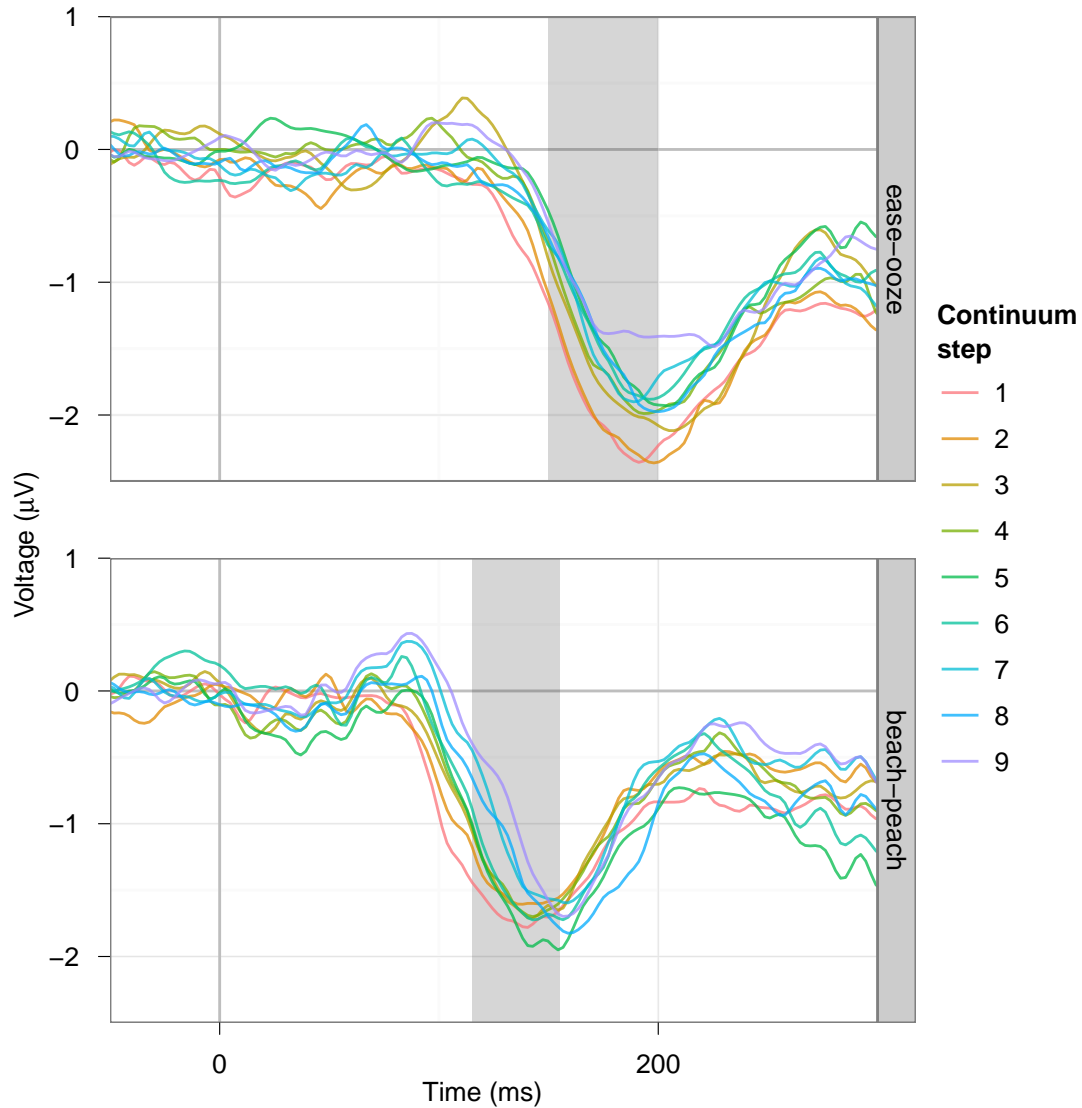
Figure 5.5: Experiment 8 results — ERP waveforms for the frontal channels by continuum step. For the *beach-peach* VOT continuum (top panel), step 1 corresponds to the /b/ endpoint and step 9 corresponds to the /p/ endpoint. N1 amplitude decreased with increasing VOT. For the *ease-ooze* formant continuum (bottom panel), step 1 corresponds to the /i/ endpoint and step 9 corresponds to the /u/ endpoint. N1 amplitude decreased with decreasing formant frequency. Shaded areas indicate time ranges used to compute mean N1 amplitude.
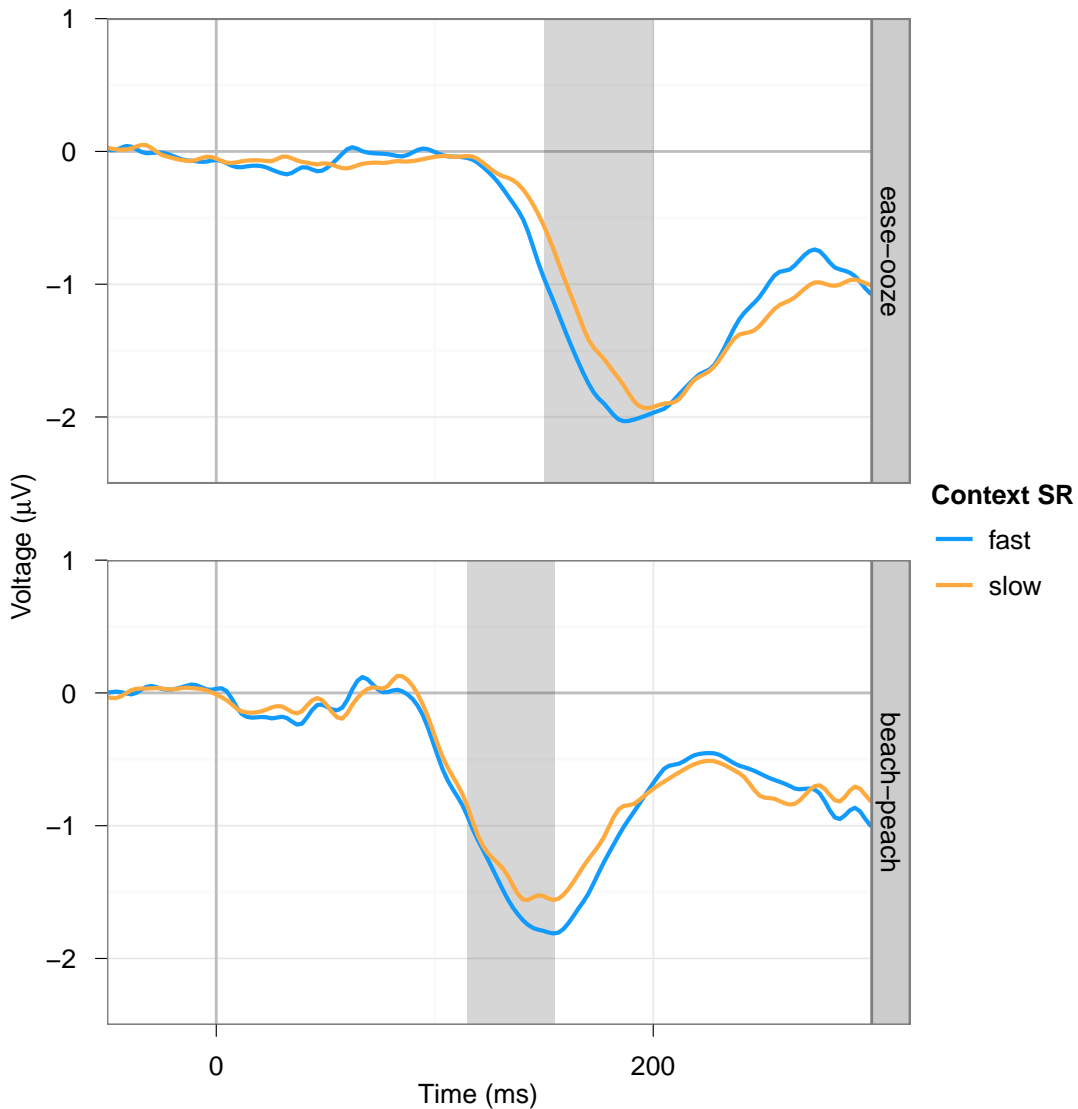
Figure 5.6: Experiment 8 ERP waveforms — ERP waveforms for the frontal channels by speaking rate. The top panel shows ERPs for the *ease-ooze* stimuli, and the bottom panel shows ERPs for the *beach-peach* stimuli. Shaded areas indicate time ranges used for mean N1 amplitude.
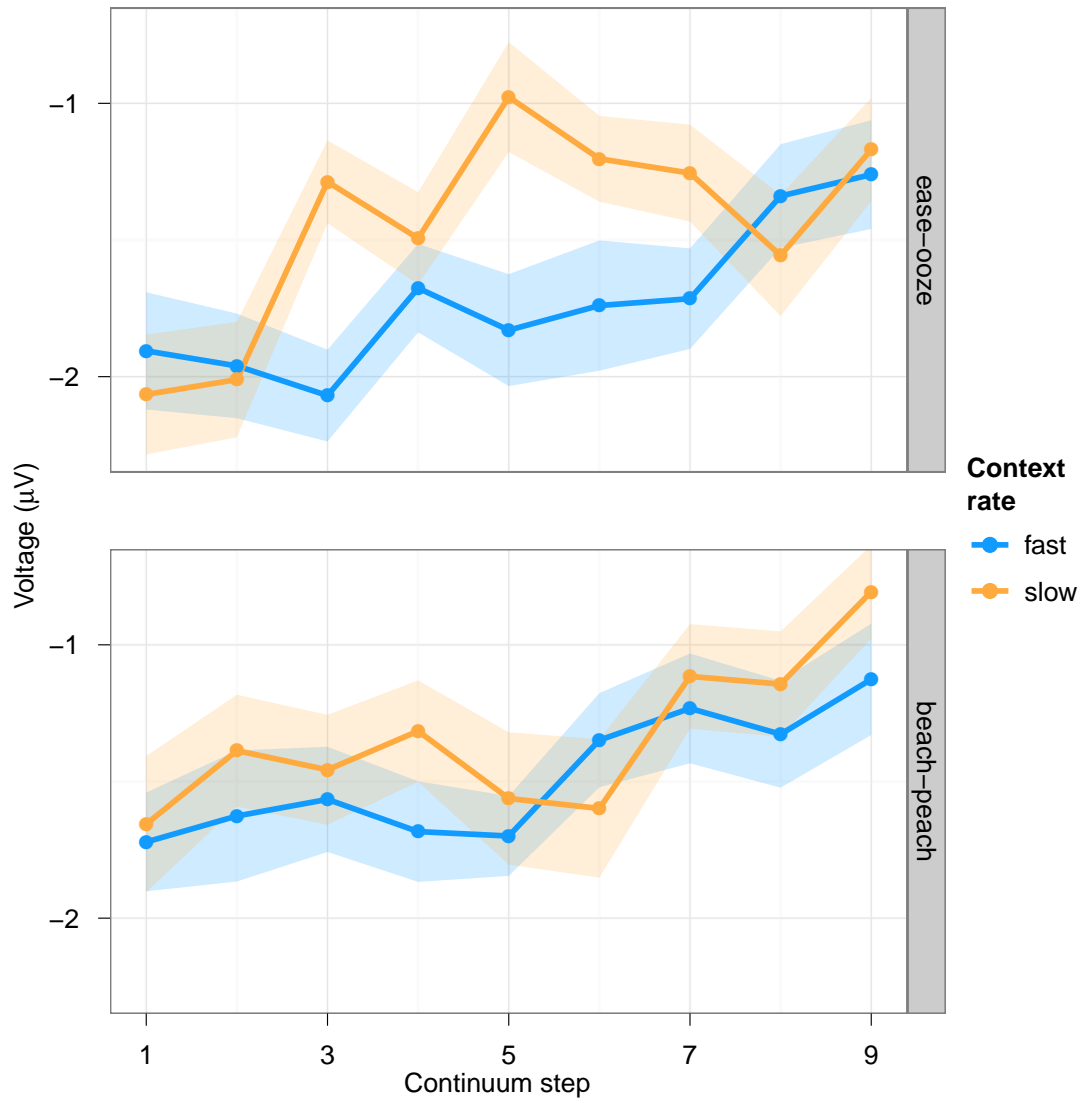
Figure 5.7: Experiment 8 results — N1 amplitude. Mean amplitude as a function of continuum step and speaking rate for the two stimulus continua (top panel: *ease-ooze* stimuli; bottom panel: *beach-peach* stimuli). Shaded areas indicate standard error.

b=-0.004, p_{MCMC}=0.886; target type x SR: b=0.16, p_{MCMC}=0.296; VOT x target type x SR: b=-0.076, p_{MCMC}=0.229).

Next, N1 amplitude was analyzed as a function of listeners' responses for trials on which they made a "target" response. Figures 5.8 and 5.9 show mean N1 amplitude for these data as a function of step and SR, respectively.

A linear mixed-effects model with VOT step, SR, and target (*beach* vs. *peach*) as fixed effects (r_{max}=-0.830[4]) did not find any significant effects (VOT: b=0.040, p_{MCMC}=0.286; target: b=-0.053, p_{MCMC}=0.774; SR: b=-0.12, p_{MCMC}=0.526; VOT x SR: b=-0.024, p_{MCMC}=0.732; target x SR: b=-0.007, p_{MCMC}=0.986; VOT x target x SR: b=-0.13, p_{MCMC}=0.424), though there was a marginal VOT x target interaction (b=0.14, p_{MCMC}=0.067). This is similar to the pattern of results observed in Experiment 4, where an effect of VOT was seen in the overall dataset but only a marginal effect was observed for the target-response trials.

The same analyses were performed to examine N1 responses to the *ease-ooze* stimuli. A linear mixed-effects model with formant frequency step, SR, and target type entered as fixed effects (r_{max}=0) showed a main effect of step (b=0.087, p_{MCMC}<0.001), with N1 amplitude decreasing with decreasing formant frequency, that is, larger N1s for the /i/ endpoint than for the /u/ endpoint. This demonstrates that variation in F2 and F3 along an /i/-/u/ vowel continuum produces linear changes in N1 amplitude, similar to the effects observed for the other phonological contrasts and acoustic cue dimensions studied. The direction of the effect is the same as the one seen for the /ɛ/-/æ/ F1 continuum in Experiment 4.

There was also a main effect of SR (b=-0.28, p_{MCMC}<0.001) with larger N1s for the fast SRs than for the slow SRs. This is the same pattern of effects seen for the *beach-peach* VOT stimuli. Given that variation in rate does not produce changes in /i/ and /u/, it seems likely that this is due to an effect other than context compensation. Moreover, it suggests that the SR effect observed for the *beach-peach* continuum may have been due to the same

---

[4]As in Experiment 6, we expect r_{max} to be high for these models because the data are grouped by listeners' responses (resulting in an unbalanced model).
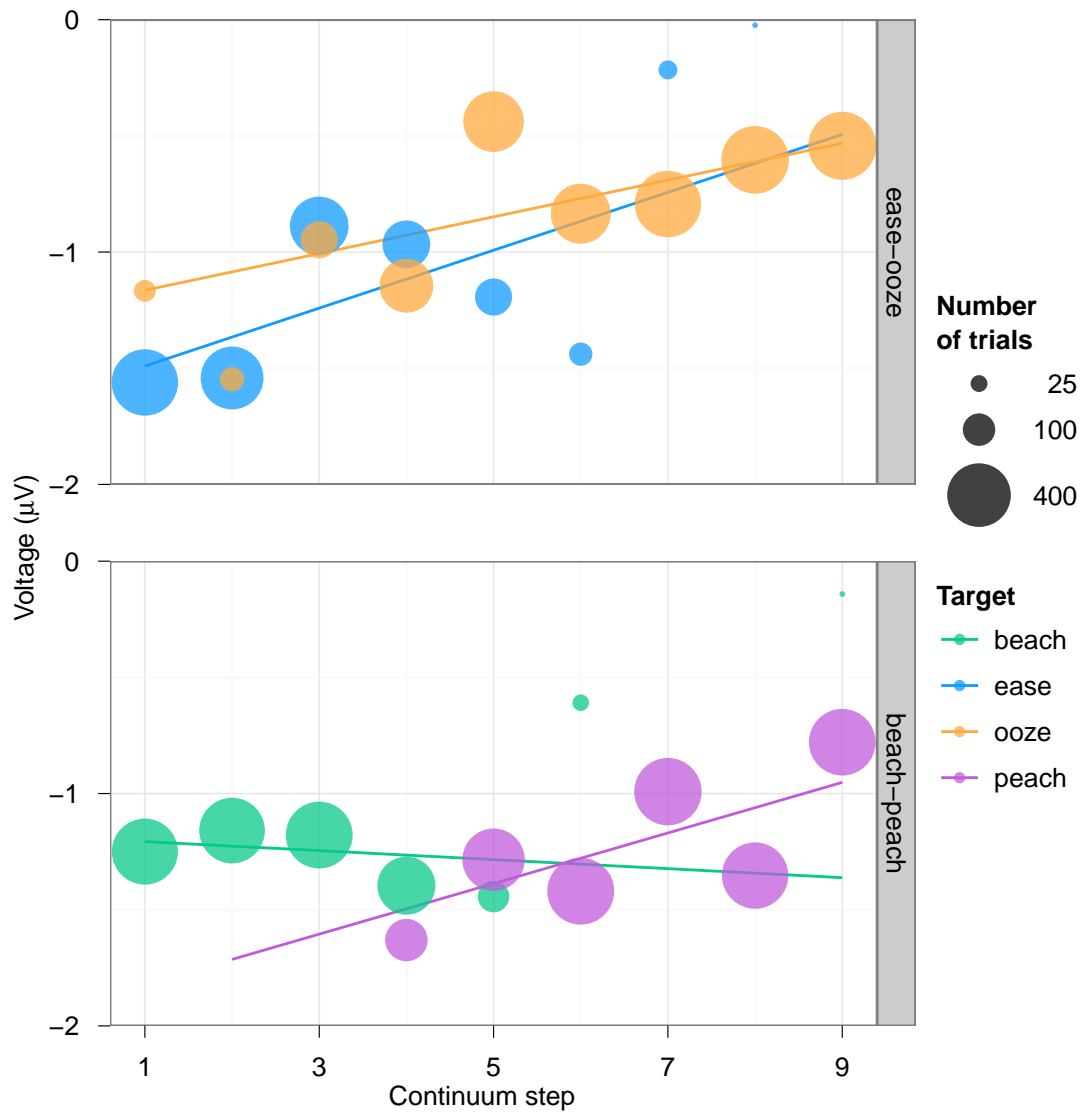
Figure 5.8: Experiment 8 results — N1 amplitude by continuum step for target-response trials. Mean amplitude for the target-response trials as a function of target word and continuum step for each stimulus continuum (top panel: *ease-ooze* stimuli; bottom panel: *beach-peach* stimuli). The size of each data point is proportional to the number of trials in that condition, and lines represent weighted linear models. The models include data from all conditions, but some data points are not plotted here because they fell outside the range of values for the other data points.
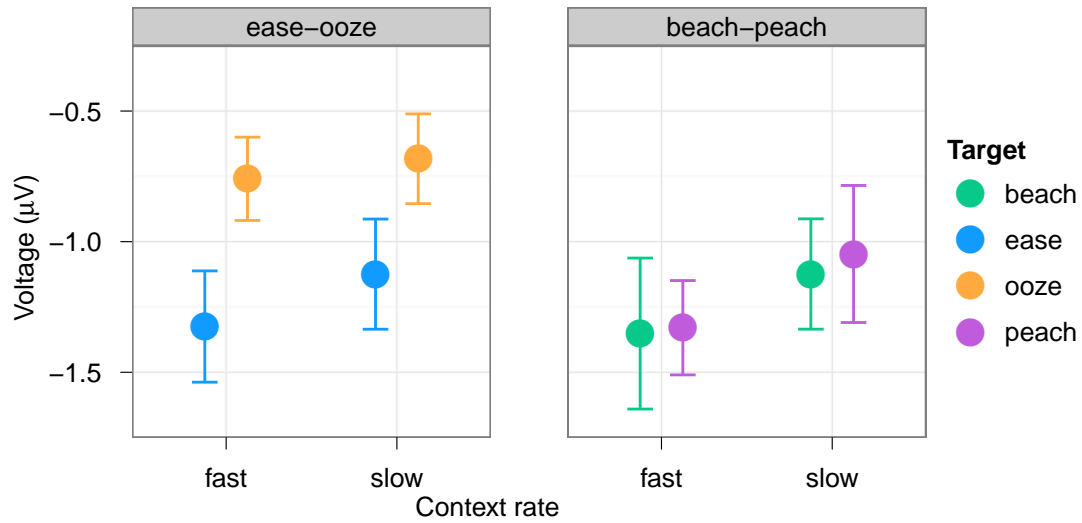
Figure 5.9: Experiment 8 results — N1 amplitude by speaking rate for target-response trials. Mean amplitude for the target-response trials as a function of target word and speaking rate. Error bars indicate standard error.

process. That is, N1s appear to be larger in the context of fast speech overall, regardless of the acoustic properties of the particular speech sounds.

Neither the effect of target type nor any of the interactions were significant (target type: b=0.023, p_{MCMC}=0.773; formant frequency x target type: b=0.045, p_{MCMC}=0.136; formant frequency x SR: b=-0.004, p_{MCMC}=0.906; target type x SR: b=0.030, p_{MCMC}=0.869; formant frequency x target type x SR: b=0.072, p_{MCMC}=0.236).

A linear mixed-effects model examining N1 amplitude on the target-response trials was run with formant frequency step, target (*ease* vs. *ooze*), and SR as fixed effects (r_{max}=-0.750). There was a main effect of formant frequency (b=0.095, p_{MCMC}=0.009) with larger N1s for higher F2 and F3 frequencies (the /i/ endpoint; the same pattern seen in the initial analysis). Other effects were nonsignificant (target: b=0.038, p_{MCMC}=0.774; SR: b=-0.012, p_{MCMC}=0.914; formant frequency x target: b=-0.12, p_{MCMC}=0.127; formant frequency x SR: b=0.085, p_{MCMC}=0.215; target x SR: b=-0.296, p_{MCMC}=0.456; formant frequency x target

x rate: b=-0.10, p_{MCMC}=0.516). These results suggest that the linear effect seen in the overall analysis was not an artifact of averaging across categorical N1 responses, providing additional evidence that N1 amplitude reflects cue encoding for formants.

5.3.2.3 *P3 amplitude*

P3 amplitude was assessed similarly to Experiment 6. Figures 5.10 shows grandaverage ERP waveforms for the parietal channels as a function of trial type (target vs. nontarget), and Figure 5.11 shows ERPs as s function of distance from the target endpoint, demonstrating that P3 amplitude is larger for target trials than for nontarget trials and decreases with distance from the target endpoint.

SR conditions were coded in a way similar to the one used for talker gender differences in Experiment 6, that is, whether the SR was consistent with the target endpoint or not. Thus, for the *beach*-target trials, slow SRs were coded as consistent (since listeners' voicing estimates should be shifted toward the voiced [target] endpoint in the context of slow speech if they are compensating for its effects) and fast SRs were coded as inconsistent (since listeners voicing estimates should be shifted toward the voiceless [nontarget] endpoint). Similarly, for the *peach*-target trials, fast SRs were coded as consistent and slow SRs were coded as inconsistent. Because context consistency is undefined for the *ease-ooze* stimuli (which do not vary as a function of rate), these effects were only examined for the *beach-peach* stimuli. Figure 5.12 shows ERPs as a function of preceding SR.

Mean P3 amplitude was analyzed by computing the mean voltage for the average of the three parietal channels from 300 to 700 ms after the onset of the target word. Figure 5.13 shows P3 amplitude for target trials as a function of target distance (for both continua) and context (for the *beach-peach* stimuli). P3 amplitude appears to decrease with distance from the target endpoint, though there does not appear to be a systematic differences between the consistent and inconsistent context conditions for the *beach-peach* stimuli.

To examine whether VOT and SR showed effects on the P3, mean amplitudes for the
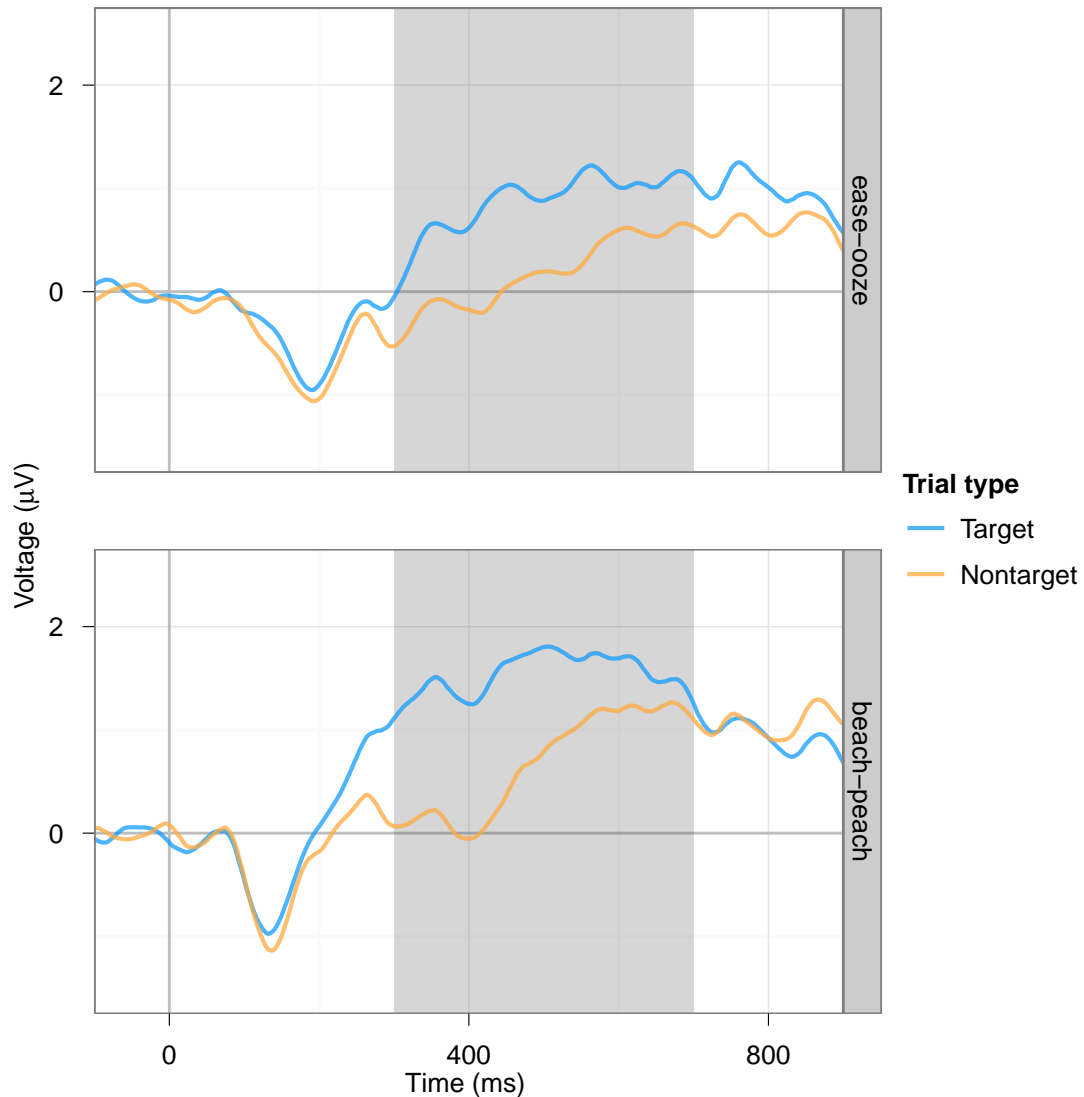
Figure 5.10: Experiment 8 results — ERP waveforms for the parietal channels by trial type. For the *ease-ooze* stimuli (top panel), the *ease* and *ooze* target blocks are the "target" trials and the *beach* and *peach* blocks are the "nontarget" trials. For the *beach-peach* stimuli (bottom panel), the *ease* and *ooze* blocks are "nontarget" trials, and the *beach* and *peach* blocks are "target" trials. P3 amplitude was larger for the target trials than the nontarget trials. Shaded areas indicate time range used to compute mean P3 amplitude.
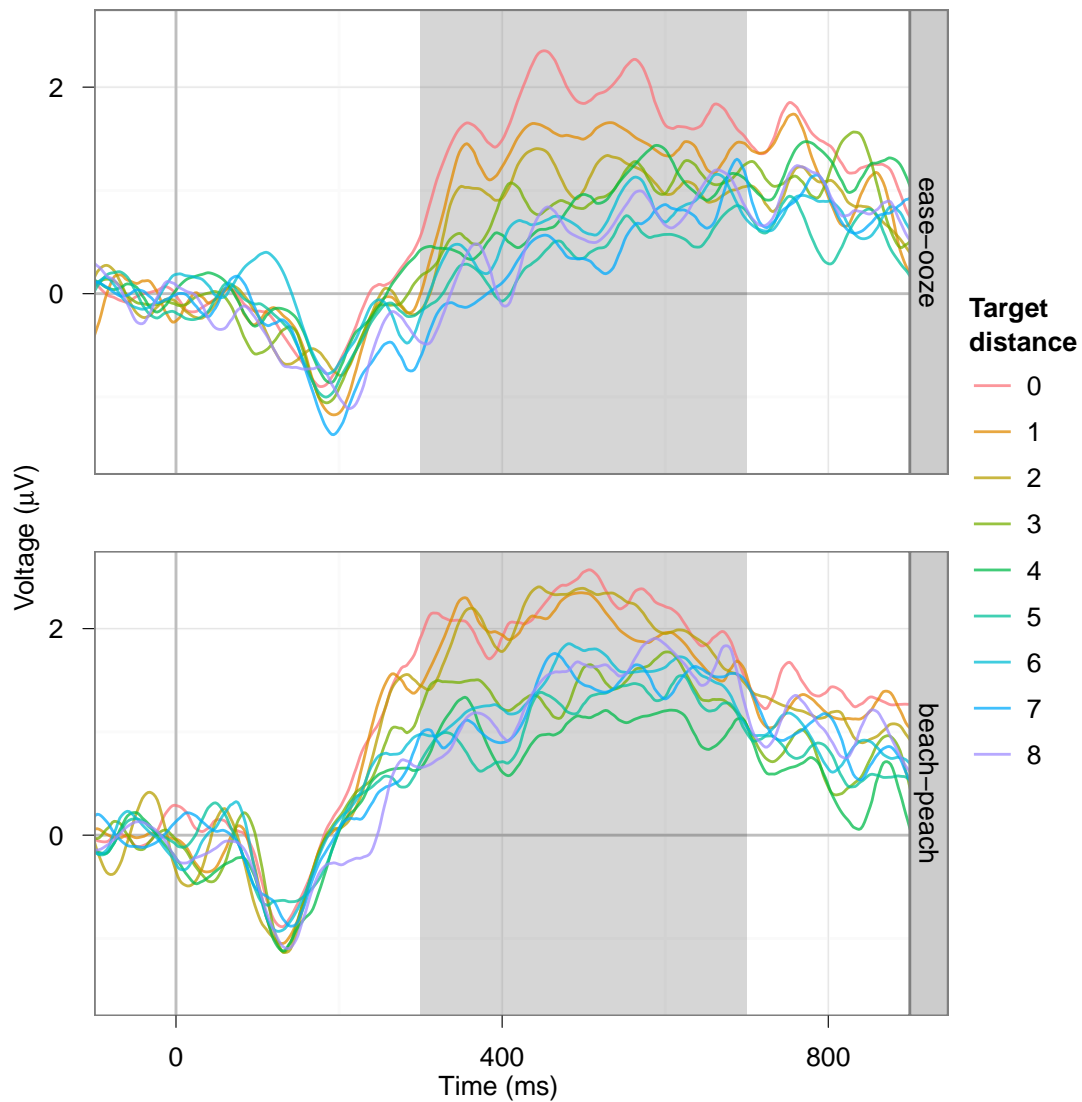
Figure 5.11: Experiment 8 results — ERP waveforms for the parietal channels by target distance. The top panel shows ERPs for the *ease-ooze* stimuli, and the bottom panel shows ERPs for the *beach-peach* stimuli. Target distance corresponds to the number of continuum steps away from the target endpoint (e.g., when *beach* is the target, step 1 along the VOT continuum corresponds to a target distance of 0, and step 9 corresponds to a target distance of 9). P3s decreased with distance from the target endpoint. Shaded areas indicate time range for P3 amplitude.
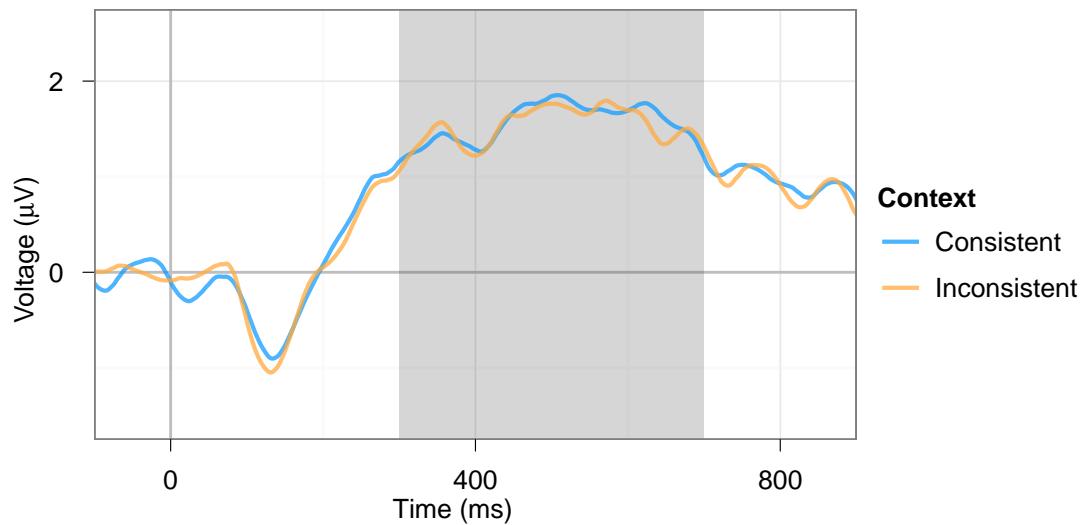
Figure 5.12: Experiment 8 results — ERP waveforms for the parietal channels by context. Preceding SR is coded as either consistent or inconsistent with the relevant target endpoint (e.g., when *beach* is the target, the slow SR condition is coded as "consistent" since it should produce more "beach" responses). Only ERPs for the *beach-peach* stimuli are shown here, since context consistency is undefined for the *ease-ooze* stimuli. Shaded areas indicate time range for P3 amplitude.

beach-peach stimuli were first analyzed using a linear mixed-effects model with target type (*beach/peach* vs. *ease/ooze*) as a fixed effect. The model showed a significant effect (b=-0.86, p_{MCMC}=0.003) with larger N1s for the target blocks than the nontarget blocks. Mean amplitude for the target blocks was then examined using a model with target distance and context consistency entered as fixed effects (r_{max}=0). The model showed a main effect of target distance (b=-0.10, p_{MCMC}<0.001) with larger N1s near the target endpoint. The effect of context consistency (b=0.029, p_{MCMC}=0.783) and interaction (b=0.020, p_{MCMC}=0.615) were not significant.

P3 amplitude was then analyzed as function of distance from listeners' VOT boundaries. Figure 5.14 shows grandaverage ERP waveforms as a function of boundary distance, and Figure 5.15 shows ERPs for the *beach-peach* stimuli as a function of whether the preceding SR was consistent with the target endpoint. Figure 5.16 shows mean P3 amplitudes
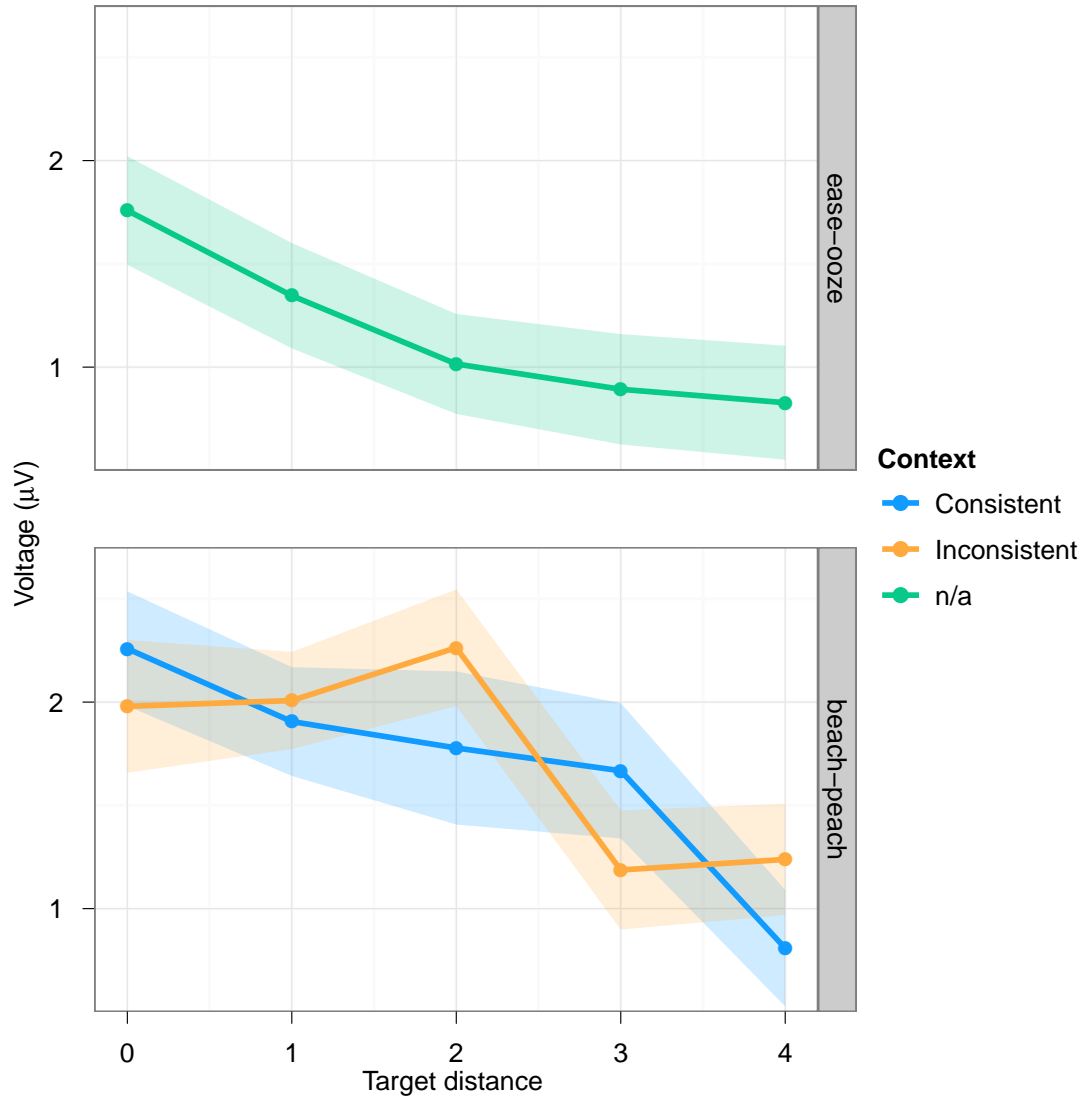
Figure 5.13: Experiment 8 results — P3 amplitude by target distance and context. Mean amplitude as a function of distance from the target endpoint and preceding context for the target trials. For the *beach-peach* stimuli, context is indicated as either consistent or inconsistent with the target endpoint. The two SR conditions are collapsed for the *ease-ooze* stimuli, since context consistency is undefined for them (labeled here as "n/a"). Shaded areas indicate standard error.

for these data. A linear mixed-effects model for the target-response trials with boundary distance, context consistency, and target (*beach* vs. *peach*) as fixed effects (r_{max}=-0.094) found a main effect of boundary distance (b=0.22, p_{MCMC}<0.001) with larger N1s near the target endpoint. None of the other main effects or interactions were significant (target: b=-0.024, p_{MCMC}=0.871; context: b=0.020, p_{MCMC}=0.884; boundary distance x target: b=0.041, p_{MCMC}=0.355; boundary distance x context: b=-0.068, p_{MCMC}=0.464; target x context: b=0.091, p_{MCMC}=0.768; boundary distance x target x context: b=-0.054, p_{MCMC}=0.796).

The same analyses were run on the *ease-ooze* data (except that the context consistency factor was not included, since it is not defined for these stimuli). An initial linear mixed-effects model with target type as a fixed effect was significant (b=-0.70, p_{MCMC}=0.006). A model for the target trials with target distance as a fixed effect also found a significant effect (b=-0.146, p_{MCMC}<0.001), showing that P3 amplitude decreased with distance from the target endpoint.

Next, target distances relative to each listeners' category boundary along the *ease-ooze* continuum were computed and a linear mixed-effects model with boundary distance and target (*ease* vs. *ooze*) for the target-response trials was run (r_{max}=-0.166). The model showed a main effect of boundary distance (b=0.142, p_{MCMC}<0.001), again with P3 amplitude decreasing with distance from the target endpoint. The effect of target (b=0.078, p_{MCMC}=0.590) and interaction (b=-0.009, p_{MCMC}=0.944) were not significant. Overall, these results are consistent with the previous experiments demonstrating linear effects of acoustic differences within individual phonological categories on P3 amplitude.

5.3.3 Discussion

These results do not provide any evidence indicating that SR has an effect on N1 amplitude to the onset of the target word. However, there was an overall effect of SR for the *ease-ooze* continuum and a marginal effect for the *beach-peach* continuum, though not in the direction predicted for compensation. Thus, while these results support raw-cue
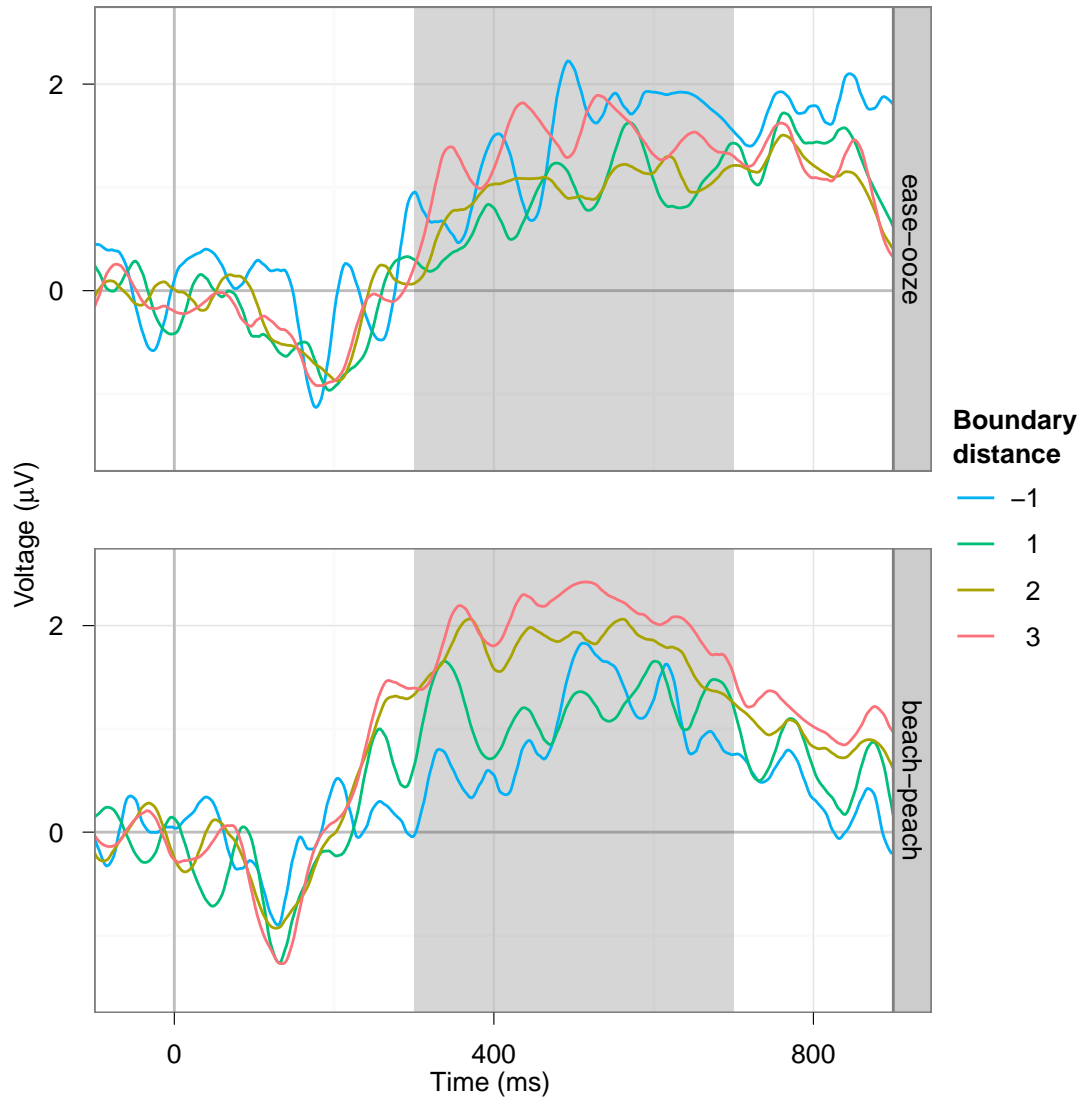
Figure 5.14: Experiment 8 results — ERP waveforms for the parietal channels by category boundary distance on target-response trials. The top panel shows ERPs for the *ease-ooze* stimuli, and the bottom panel shows ERPs for the *beach-peach* stimuli. Higher step numbers indicate points closer to the target endpoint relative to each participants' category boundary. Shaded areas indicate time range used to compute mean P3 amplitude.
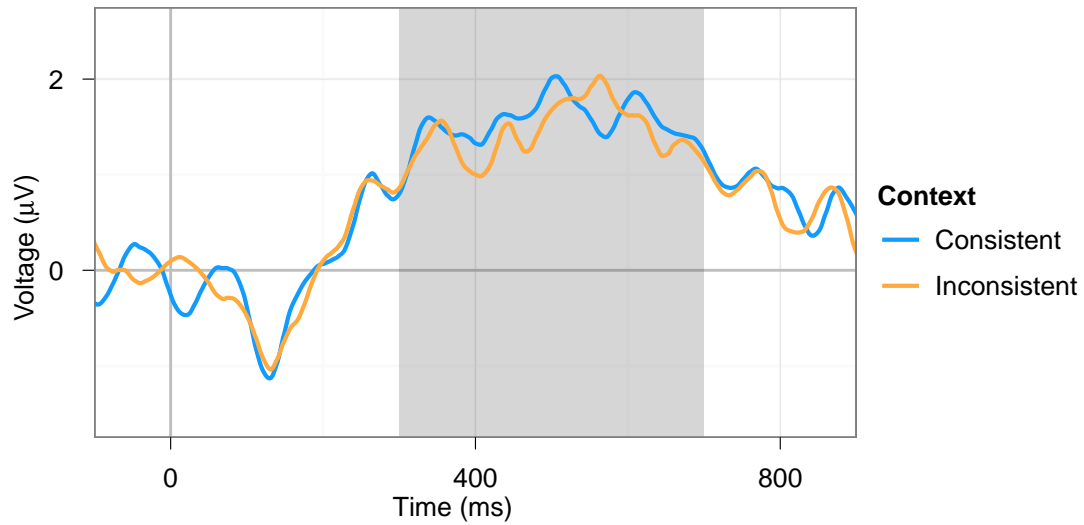
Figure 5.15: Experiment 8 results — ERP waveforms for the parietal channels by context on target-response trials. Context conditions are plotted as a function of whether the they were consistent or inconsistent with the target endpoint. Only data from the *beach-peach* continuum are shown here, since context consistency is undefined for the *ease-ooze* stimuli. Data are averaged across the four rStep conditions shown in Figure 5.14 (-1, 1, 2, and 3). Shaded area indicates time range for P3 amplitude.

encoding models, which predict no rate effects during cue encoding, this conclusion must be considered in light of other results. Indirectly, these results also suggest a compensation model that uses extrinsic encoding via feedback. Here, since category-level information about context is not available, it may be that the information needed to encode cues relative to context is not present.

The overall effect of SR, with fast rates producing larger N1s regardless of the acoustic cues in the target word, could be indicative of increased work being done during cue encoding for fast speech. This would be a relatively straightforward explanation if fast speech was also more difficult to understand, however, speech intelligibility is not affected by speaking rate (Bradlow et al., 1995). Thus, this may reflect some other process active during cue encoding.

The behavioral results also did not show any evidence of compensation for rate
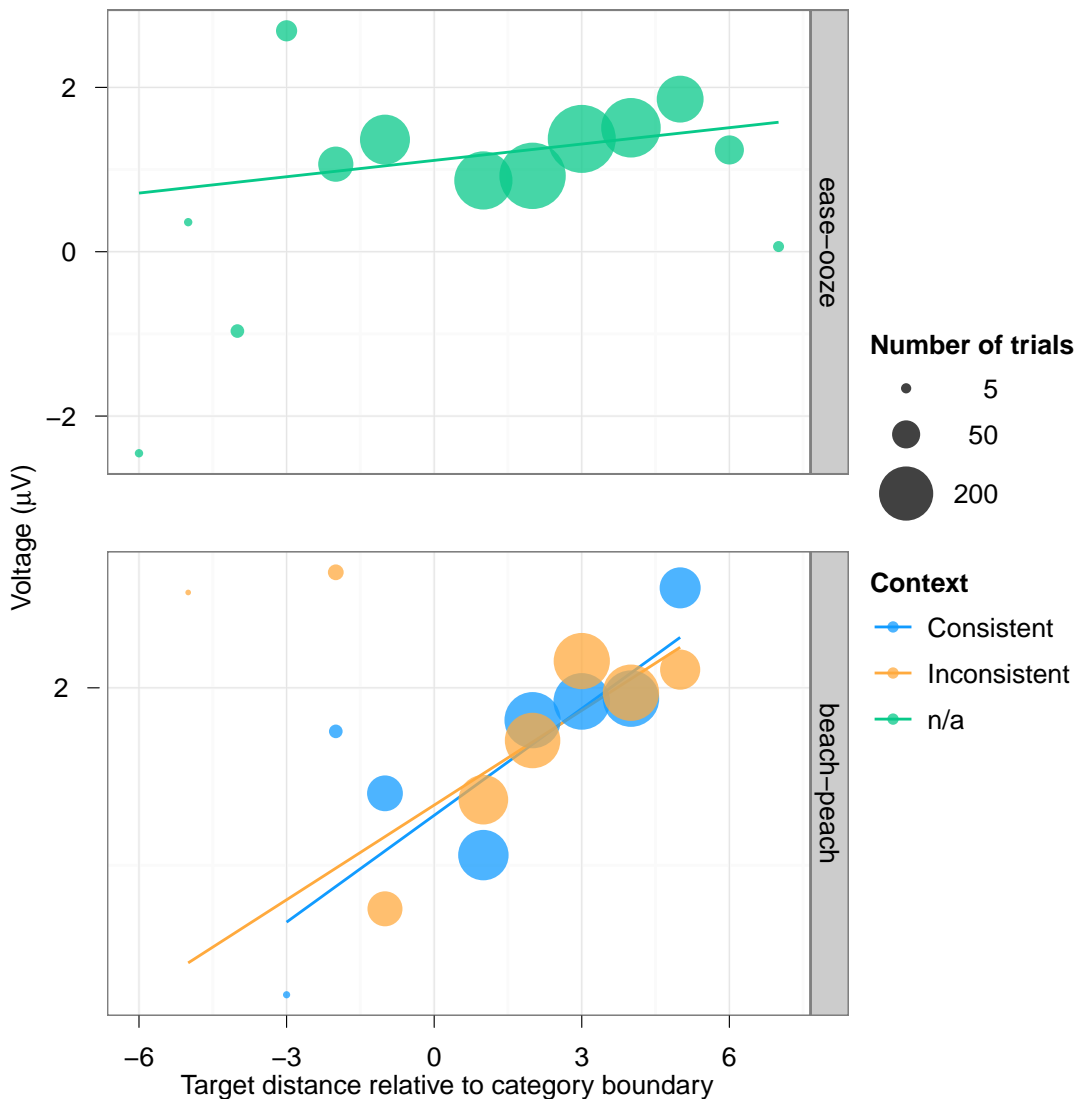
Figure 5.16: Experiment 8 results — P3 amplitude by category boundary distance and context. Mean amplitudes for the target-response trials as a function of distance from participants' category boundaries and, for the *beach-peach* stimuli (bottom panel), whether the preceding SR was consistent with the target endpoint (for the *ease-ooze* stimuli [top panel], the average of the two SR conditions is shown). The size of each data point is proportional to the number of trials in that condition, and lines represent weighted linear models.

variability. This is potentially problematic for the interpretation that rate does not have an effect on cue encoding — it may be that it has no effect at all on listeners' voicing judgments. However, previous work (Summerfield, 1981; Wayland et al., 1994), as well as Experiment 7, suggest that SR does has an effect on voicing judgments. There are several differences between Experiment 7 and the present experiment that could have contributed to the lack of an effect in this experiment. The most likely candidates are that (1) the new set of stimuli used in this experiment, and (2) the target detection task that used instead of a 2- or 4AFC task. Given that the range of SR values used is similar for the two experiments, it seems that the task may have masked effects of SR.

To test this, several participants were run in a 2AFC task with the *beach-peach* stimuli from this experiment. The experimental procedure was the same as in Experiment 5 — listeners indicated whether the word the heard at the end of the sentence was *beach* or *peach*. Figure 5.17 shows listeners responses in this task as a function of VOT step and SR. The results confirmed that an effect of SR, approximately the same size as that seen in Experiment 7, can be seen with these stimuli (b=-0.51, z=-2.32, p=0.020, r_{max}=-0.091). Listeners made more "peach" responses in the context of a fast carrier phrase, and more "beach" responses in the context of a slow carrier. Thus, it appears that the target detection task simply masked the effects of SR in the present experiment.

The pattern of N1 differences for VOT at the continuum endpoints (i.e., shorter VOTs produce larger N1s) in previous experiments was similar for both 2AFC (Experiment 2) and target detection (Experiment 4) tasks. Thus, it does not seem that the task would affect listeners' N1 responses to the two SR conditions either, though indexing functions for ERP components may be specific to the task used (Mordkoff & Grosjean, 2001). If N1 responses are the same across tasks, the overall results are best explained by raw-cue encoding models in which SR serves as a weak phonetic cue (Toscano & McMurray, 2010) or influences the mapping between cues and categories (Smits, 2001b) via feedforward processes.
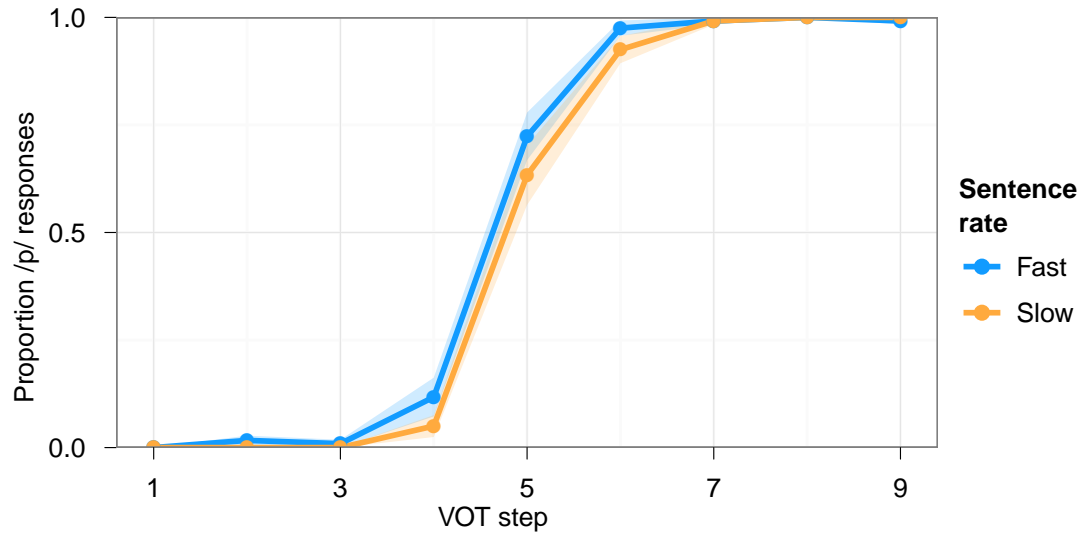
Figure 5.17: Experiment 8 — results of 2AFC follow-up task. Listeners' categorization responses for the stimuli from Experiment 8 but in a 2AFC task, showing that preceding rate has an effect on their voicing judgments.

5.4 General discussion

Along with the results of the experiments on talker variability, the results of this set of experiments suggest that relative cue encoding is mediated by feedback from category-level representations, but that continuous information from context can still influences listeners' decisions. This combines aspects of feedback-driven, extrinsic encoding models (e.g., C-CuRE) with feedforward, raw-cue encoding models (e.g., HICAT, WGMM). In fact, C-CuRE uses both raw and relative cue encoding to handle contextual variability. However, the way raw cues are used, by directly providing information about different phonological categories, seems unusual for effects like SR. Using SR in this way would suggest that listeners simply activate voiced categories more in the context of slow speech and voiceless categories more in the context of fast speech. There are ways to measure SR other than the global speaking rate that may be more sensible under this approach. For example, listeners could use VL information from the preceding sentence to bias phonological categories. This

seems more reasonable, and it fits with the results showing that VL acts as a weak phonetic cue (McMurray, Clayards, et al., 2008; Toscano & McMurray, 2011a) and those showing that rate information occurring close in time to the stop consonant has larger effects (Repp & Lin, 1991; Summerfield, 1981).

The lack of an effect here also suggests that, when cue-level context effects are observed, they are driven by category-level information. Indeed, a feedback model that uses information from higher-level, but continuous, estimates of rate could have shown an effect of SR on VOT encoding in Experiment 8. This would have been indistinguishable from a lateral model in this case, and both types of models are logically quite similar. However, given that compensatory effects of rate on VOT encoding were not found, both a lateral and a continuous-information feedback account seem implausible. Instead, these results suggest that, if listeners do use feedback to compute relative cues, they do so based on feedback from specific categories.

The two secondary goals of these experiments also support the conclusions above. First, we observed a single N1 peak at both ends of the voicing continuum (seen here in Experiment 8, as well as in Experiments 2 from Chapter 3). This provides additional evidence that listeners process continuous changes in VOT. Second, effects of preceding SR were observed in natural speech (Experiment 7 and the follow-up to Experiment 8). This supports previous results and provides evidence against purely intrinsic encoding. Although this effect wasn't observed in the target detection task used in Experiment 8, this seems to be the result of the task used rather than the stimuli, as shown by the presence of an effect in the follow-up to Experiment 8.

Together, the results of the talker and rate variability experiments help to distinguish between different models of speech perception that have been proposed to deal with contextual variability. We next ask how listeners compensate for the third main source of variance in speech, coarticulation. The final experiment addresses this.

CHAPTER 6
COARTICULATORY CONTEXT

6.1   Background

The final experiment examines the effects of coarticulatory context. The fundamental issue here is that the acoustic-phonetic instantiation of a phoneme is never independent of neighboring phonemes. As in the previous chapters on talker and rate compensation, the two principles that distinguish models of context compensation (type of encoding and direction of information flow) also apply here, and several models that vary along these dimensions have been proposed. However, one difference between coarticulatory effects and the effects we have looked at previously is that coarticulation potentially provides two sources of anticipatory context information that listeners can take advantage of. The first corresponds directly to the type of contextual variability discussed for talker and rate differences: the effect of the preceding context can alter the acoustic properties of subsequent cues, leading to ambiguity and difficulty assigning the correct phonological category. Listeners could handle these effects in the same ways they handle talker and rate effects, by compensating for differences due to preceding phonemes at the level of cue encoding (i.e, using relative cue values) or categorization (using raw cues).

However, unlike talker and rate effects, coarticulation about the identity of subsequent phonemes is also present in the preceding context. For example, the syllable /si/, consists of a period of frication followed by a period of vocalic energy. Information about the fricative will carry over into the vocalic portion (creating a context effect on the vowel similar to the way talker and rate can affect the acoustic form of the vowel). In addition, information about the vowel is present during the frication which can lead to modifications of the same cues that identify the fricative (e.g., an /s/ can be more /ʃ/-like in some contexts). These modifications in the frication can serve as an additional source of anticipatory information about vowel identity, or they could alter fricative categorization. Critically, if

listeners treat these frication differences as anticipatory cues to the vowel (which they are), listeners may not compensate for differences in the vowel due to the fricative, as they may do for other sources of context information in vowel identification. That is, because this information directly cues vowel categories, we would not expect listeners to factor it out of their estimates of cue-values for the subsequent vocalic portion. If, however, they treat vowel information in the frication as contextual variability (i.e., variation in the talker's production of the fricative), we would expect compensation of a form similar to that seen for other sources of contextual variability.

Here, Repp's (1982) distinction between trading relations and context effects becomes apparent. The effect of preceding information caused by other phonemes (e.g., whether the fricative is an /s/ or /z/ in a fricative-vowel syllable) on subsequent acoustic cues (information in the vocalic portion) would be considered a context effect. In contrast, coarticulatory information that is present in one segment, but attributable to another (e.g., differences during the frication between the syllables /si/ and /su/), would be a source of information for a trading relation between the cues to the relevant phonological distinction (/i/ vs. /u/). However, there are multiple ways this information can have an effect on listeners cue encoding (if it has any effect at all). Since information in the frication could be treated as either an additional cue (information about the vowel) or as an effect of contextual differences (information about the fricative) we would expect different effects on listeners' encoding of subsequent vocalic information.

If listeners treat this coarticulatory information as a cue to the vowel, we would expect that they use it to either facilitate recognition of the vowel or prime activation of cue-values associated with a particular vowel category. This would be similar to the types of effects seen in audiovisual speech perception (van Wassenhove et al., 2005), where N1 responses have shorter latencies in the context of audiovisual speech (which provides similar anticipatory cues) than in the context of audio-only speech.

Alternatively, if listeners treat this information as a source of contextual variability (i.e., differences in how the talker produced an /s/ or information indicating different fricatives), we would expect effects on cue encoding that are similar to other types of context effects. For example, if the coarticulation in the frication causes it to be identified (even partially) as a different phoneme, this may cause listeners to treat the information in the vowel differently.

Thus, whether listeners treat anticipatory coarticulation as an additional cue (about the vowel) or as context information (about the fricative) can be determined by looking at how they encode cues in different coarticulatory contexts.

This was examined in a single, somewhat exploratory experiment that used a slightly different design than the prior studies. Before presenting the experiment, I will review the phonetic and perceptual data on coarticulatory effects, with a particular emphasis on fricative-vowel coarticulation, which is examined in the present experiment. In addition, evidence for and against proposed models of coarticulatory compensation, and how they may differ for trading relations and true context effects, are discussed.

6.1.1   Phonetic data

A large number of phonetic studies have examined coarticulatory effects of different phonological contrasts, including those that affect vowels (House & Fairbanks, 1953; Stevens, House, & Paul, 1966; Magen, 1997; Cole et al., 2010), stop consonants (Öhman, 1966; Kewley-Port, 1982; Repp, 1982), and fricatives (Fujisaki & Kunisaki, 1978; Soli, 1981). Overall, coarticulation has similar consequences across different phonological contrasts, such that the acoustic properties of the first segment reflect properties of how the upcoming and prior segments are produced (and vice versa).

Here, I will focus on coarticulatory effects of fricative-vowel pairs, since these two classes of sounds will be examined in the present experiment. Fricative-vowel syllables are characterized by a segment of aperiodic energy (frication; which may also contain periodic

energy, particularly for voiced fricatives) followed by a segment of periodic energy (the vocalic portion). Both fricatives and vowels are distinguished by spectral cues. As discussed in Chapter 4 F1 and F2 value serve as primary cues to vowel quality (Peterson & Barney, 1952; Hillenbrand et al., 1995). Formant frequencies at vowel onset also distinguish fricative place of articulation (Jongman et al., 2000), but there are substantial spectral changes within the frication that are much more robust cues to place of articulation (Forrest, Weismer, Milenkovic, & Dougall, 1988; Jongman et al., 2000). These are usually captured in terms of broad spectral measures, such as the overall spectral mean and variance.

Differences in fricative place of articulation and voicing also exert coarticulatory effects on several types of surrounding segments, including vowels. House and Fairbanks (1953) examined the coarticulatory effects of a large number of consonants, including fricatives, on vowel sounds. They found that the preceding consonant has effects on vowel duration, fundamental frequency, and relative intensity. Similarly, Stevens, Blumstein, Glicksman, Burton, and Kurowski (1992) found that differences in fricative voicing affect F1 transitions into subsequent vowels. Repp (1982) reported similar coarticulatory effects of fricatives for combinations of fricatives (/s,ʃ/), stops (/t,k/), and vowels (/a,u/) in FCV and VFCV contexts. They found that formant values, which provide acoustic cues to stop place of articulation as well as vowel quality, were higher at the onset of the stop after /s/ than after /ʃ/, consistent with the formant differences for those two fricatives and demonstrating a coarticulatory effect of the fricative. These results demonstrate true context effects of preceding fricatives on subsequent vowels. Crucially, they also suggest that in order to accurately identify the vowel, a listener would need to compensate for differences due to fricative context.

Vowels also create coarticulatory effects on neighboring phonemes, including stops (Stevens et al., 1966) and vowels in adjacent segments (Cole et al., 2010), and several studies have examined the acoustic properties of fricatives when followed by different vocalic con-

texts (Hughes & Halle, 1956; Heinz, 1961; Fujisaki & Kunisaki, 1978; Soli, 1981). The quality of the subsequent vowel, in particular, whether the vowel is rounded or not, has been shown to have large effects on preceding fricatives (Daniloff & Moll, 1974). Fujisaki and Kunisaki (1978) used the analysis-by-synthesis approach (Lewis, 1936)[1] to develop a model of fricative production that captured differences in the acoustic properties of Japanese /s/ and /ʃ/ depending on vowel context. They found that energy in the frequency region of F1 during frication, one of several cues distinguishing these fricatives, is influenced by the following vowel (e.g., lower for rounded vowels than for unrounded vowels). Soli (1981) found similar effects of lip rounding, as well as an independent effect of vowel backness in a set of phonetic data for /s/, /z/, /ʃ/, and /ʒ/ spoken in the context of /a/, /i/, and /u/ vowels.

Overall, the phonetic data demonstrate that there is considerable information during the frication about upcoming vowels in FV segments and, simultaneously, considerable changes in vowel formants that are due to the frication. This is particularly true for the distinction between rounded and unrounded vowels. As a result, listeners can use information in the frication to predict the category of the upcoming vowel, treating it as an additional cue to the identity of the vowel. However, they could also treat it as information about the fricative context. Differences in lip rounding for vowels produce effects in the frication that are similar to the effects of fricative place of articulation (e.g., /ʃ/ vs. /s/). Rounded vowels have a lower spectral mean during the frication than unrounded vowels, and similarly, /ʃ/ has a lower spectral mean than /s/ (Jongman et al., 2000). As a result, an /s/ that has undergone anticipatory lip rounding can be ambiguous between /s/ and /ʃ/. Given that listeners compensate for coarticulatory differences due to fricative place, they might also compensate for the lower spectral mean in the frication caused by a rounded vowel by treating the subsequent vocalic portion as having a higher frequency. The phonetic data alone, however, do not tell us whether one, both, or neither of these processes affects cue encoding.

---

[1]Don Lewis of the Iowa Psychology Laboratory was one of the first to use the analysis-by-synthesis technique, describing it in an early report on vowel sounds produced by singers.

6.1.2 Perceptual data

There have been a large number of perceptual studies showing that listeners are sensitive to and compensate for coarticulatory effects (Liberman et al., 1952; Mann & Repp, 1981, 1980; Mann, 1980; Whalen, 1981; Yeni-Komshian & Soli, 1981; Repp & Mann, 1981). In terms of the contrasts used in the present experiment, several previous studies have found effects of fricative-vowel coarticulation on listeners' perceptual judgments.

Mann and Repp (1981) investigated this by presenting listeners with stimuli consisting of CV syllables with a frication and vocalic portion. Frication varied along a formant continuum from /s/ to /ʃ/. The formant transitions into the vocalic portion were consistent with the vowels /a/ and /u/, spoken in the context of each of the fricatives. Listeners heard each combination of frication step and vocalic segment and were asked to make judgments about the identity of the fricative. Along with the overall effect of formant frequency during the frication, there were large effects of whether the vocalic segment matched an /s/ or /ʃ/, demonstrating that listeners are sensitive to the effects of fricative information in the vocalic portion. There were also effects of the vowel, such that unrounded vowels (e.g., /a/), which have higher formant frequencies, produced more /ʃ/ responses than rounded vowels (/u/), which have lower formant frequencies. Similar effects were observed by Johnson (1991) for these vowels and by Whalen (1981) for the vowels /i/ (unrounded) and /u/ (rounded). These results are consistent with the idea that listeners are compensating for differences in the upcoming vowel by treating a fricative as having a lower frequency in the context of a vowel with a higher frequency. However, it is not clear what sort of compensation model listeners use to handle this, and these results are consistent with both cue- and category-level compensation.

Effects on listeners' vowel judgments as a function of the preceding fricative have also been observed. Ostreicher and Sharf (1967) found that listeners are above-chance at identifying vowels when given only the frication portion of a fricative-vowel segment, in-

dicating that they are able to use some anticipatory vowel information in the frication. Similarly, Yeni-Komshian and Soli (1981) examined listeners identification of vowel sounds based on information from a preceding frication segment. They recorded utterances of several fricative-vowel pairs (/s,z,ʃ,ʒ/ and /a,i,u/), and isolated the frication portion of each recording. Listeners' were presented with these stimuli and asked to identify the vowel (/a/, /i/, or /u/) that the frication came from. They were above chance at identifying the vowels, though /i/ and /u/ were more accurately identified than /a/. In addition, listeners' accuracy varied as a function of the vowel the fricative was coarticulated with (an anticipatory vowel cue) and interacted with the place of articulation of the fricative (a standard context effect).

These studies indicate that listeners are perceptually sensitive to both types of contextual variability introduced by coarticulation. They show compensation for variability caused by the identity of the phoneme in the preceding segment (a context effect; Yeni-Komshian & Soli, 1981), and they show effects of anticipatory cues to the later phoneme that are present in the earlier sound, as evidenced by fricative judgments (Mann & Repp, 1981). Given this, we can now ask how different context compensation mechanisms would account for these results and what predictions they make about how listeners encode the later vocalic information in FV syllables.

6.1.3 Mechanisms for handling coarticulatory variability

Several models of context compensation have been proposed to deal specifically with coarticulatory effects. Notably, Fowler's (Fowler, 1984, 1986, 2006) gestural parsing account provides a general approach for handling coarticulation. Fowler observes that speech segments overlap in time rather than as temporally discrete units. Given this, she argues that listeners perceive speech in terms of overlapping articulatory gestures and assign particular patterns of acoustic information to the appropriate gesture in order to handle contextual variability. For example, to correctly recognize fricatives that are coarticulated with subsequent vowels, listeners assign variation in the frication that is due to the following vowel

to the vowel. Removing this variability allows the listener to more accurately identify the fricative (and, subsequently, the vowel).

Other researchers have challenged this account, arguing for auditory-based models of coarticulation compensation (Lotto & Kluender, 1998; Diehl, Lotto, & Holt, 2004; Kluender, 2003; Holt, 2005). Under these approaches, coarticulatory effects are explained by auditory contrasts (e.g., spectral contrasts) between the first and second segment. That is, in the context of a preceding high-frequency sound, a given sound is perceived as having a lower frequency than it actually does, and vice versa. To illustrate this idea, Lotto and Kluender (1998) describe a study by Mann (1980) in which she demonstrated that listeners compensate for F3 coarticulation from liquids (/l,r/) such that they are more likely to perceive a stop consonant as /ga/ when it is preceded by /al/, and as /da/ when preceded by /ar/. (Lotto & Kluender, 1998) demonstrated that similar effects are obtained when the preceding context is a nonspeech sound with energy in the frequency range corresponding to /al/ or /ar/. They argue that these results do not fit with a gestural parsing account in which listeners are attempting to recover the configuration of the talker's articulators (since there is no talker and no phoneme information for nonspeech sounds).

Recently, however, Viswanathan, Magnuson, and Fowler (2010) presented evidence against the general auditory hypothesis by examining a coarticulatory contrast that does not demonstrate spectral contrast. Listeners still showed compensation for coarticulation, and Viswanathan et al., argued that this result supports a gestural account over a general auditory account. Although the distinctions between these two approaches are critical to overall theories of speech perception, for the purposes of this experiment, both accounts would be characterized similarly along the two main dimensions used to classify the models, though, as noted in Chapter 2, it is difficult to compare gestural parsing accounts with the other models. The auditory contrast account would be classified as a lateral, extrinsic encoding model, since it handles coarticulation via interactions at a single level of processing

based on information from adjacent segments.

A feedback mechanism is also suggested by C-CuRE. Cole et al. (2010) used C-CuRE to capture effects of vowel-to-vowel coarticulation, a phenomenon in which coarticulatory information from an upcoming vowel is present in a preceding vowel (Cho, 2004; Öhman, 1966). They collected phonetic data from a number of talkers and measured formant frequencies for vowels in a variety of consonant and vowel contexts. They found that by factoring out expected acoustic variation based on coarticulatory context (the consonant and subsequent vowel), as well as information about the talker, they could greatly reduce the amount of overlap between between vowel categories in F1xF2 space. Using this approach, they were able to account for 95% of the variation in F1 and F2 values for these vowels. Thus, this provides a potentially powerful approach for handling coarticulatory effects. Like gestural parsing and auditory contrast accounts, C-CuRE uses extrinsic information (e.g., talker identity and the identity of neighboring phonemes) to encode cue-values relative to context (see also McMurray & Jongman, 2011, for a demonstration of this in fricatives). However, C-CuRE differs from the auditory contrast account in that it effectively uses category-level information, rather than cue-cue interactions, to adjust cue-values. This places it as a feedback, extrinsic encoding model in the classification system used here.

Another line of research has examined the question of what units (e.g., phonemes vs. syllables) are used to perceive coarticulated speech, which has led to debates between several raw-cue encoding models designed to account for coarticulatory effects. Mermelstein (1978) presented data suggesting that listeners do not use coarticulatory information from adjacent phonemes in a compensatory way. That is, they do not identify the preceding segment (that creates the coarticulation) as a precursor to identifying the target segment (that undergoes coarticulation). He found that when adjacent consonants and vowels were signaled by the same information (duration of the vocalic portion), listeners made independent judgments about the identity of the consonant and vowel, using the duration information

to make decisions about both and suggesting that they do not compensate for duration information attributed to one phoneme in their estimates of the other. Whalen (1989) attempted to replicate Mermelstein's (1978) study with more power and experimental control and found evidence in favor an interaction between consonant and vowel judgments. In addition, he extended these results to examine FV syllables (/si/, /ʃi/, /su/, and /ʃu/) signaled by differences in the frication (similar to the acoustic differences used in the present experiment) and found similar results.

Whalen's results contradict Mermelstein's earlier claims and suggest that listeners compensate for coarticulatory effects. However, Nearey (1990) analyzed Whalen's results using a NAPP model and found that a diphone bias, which is applied at later decision stages, can produce the same effects, even when the fricatives and vowel are categorized independently. Smits (2001a) also examined Whalen's results using HICAT, and suggested that it could better account for the compensatory effects observed using a scheme in which decisions about the preceding segment condition judgments on the target segment. In terms of the distinctions being made here, the results from both models suggest that feedforward, raw-cue encoding approaches are sufficient for explaining compensation for coarticulation. Overall though, the most significant difference between these approaches and others (e.g., auditory contrast, C-CuRE) is that they predict that initial cue encoding is not affected by coarticulatory context. Rather, any compensatory effects of coarticulation are handled at categorization or later stages.

Each of these accounts also allows for the possibility that anticipatory information about upcoming phonemes can affect listener's phonological judgments in feedforward way. However, their predictions differ depending on whether listeners treat this information as a cue to the vowel or as contextual variation related to the fricative (or both). If listeners use it as a vowel cue, both types of models predict there will be no effect on listeners' encoding of later vowel information (since they are not compensating for differences in

fricative context). If, however, listeners treat this information as a sources of contextual variability (i.e., a cue to the fricative), the two types of models make different predictions about how listeners encode later vowel information. Raw-cue encoding models predict that this information would not affect encoding of the subsequent vocalic information, but would be taken into account at a later stage of processing. In contrast, relative encoding models suggest that listeners compensate for vowel-derived differences in frication at the level of cue encoding, just as they would compensate for effects of talker, rate, or true contextual differences in the fricative produced by the talker.

6.1.4   Experiment overview and predictions

Following the logic of the previous experiments, the current experiment examined listeners' responses to a vowel sound following context information, in this case, a preceding fricative. The context compensation models described above make different predictions about how listeners encode vowel cues in FV segments and what effects we should see on N1 responses to the onset of the vocalic period. Consider the vowels /i/ and /u/ spoken in the context of an /s/ such that there is coarticulatory information about the vowel during the frication. Because /i/ has higher frequency F2 and F3 values than /u/ (Hillenbrand et al., 1995), spectral information in the /s/ will be higher frequency and unambiguously /s/-like. Therefore, /s/ sounds produced in the context of an /i/ will be referred to using the notation /s_H/ (indicating a *H*igher frequency /s/). Similarly, since /u/ coarticulation leads to lower frequencies in the frication (producing a partially /ʃ/-like sound), /s/ sounds produced in the context of an /u/ will be referred to as /s_L/ (*L*ower frequency).

This produces differences in frication frequency similar to those produced by the differences between /s/ and /ʃ/. Thus, this leads to a situation in which listeners can use the coarticulatory information in the frication as either a vowel cue (i.e., /s_L/ indicates an /u/ is arriving next) or as fricative place information (/s_L/ is more consistent with an /ʃ/). In addition, by cross-splicing the sounds, we can create mismatching stimuli in which /s_H/ is

followed by /u/ or /s_L/ is followed by /i/.

There are three possible effects that such anticipatory coarticulation can have on subsequent cue encoding. The first possibility is that the information about the vowel in the frication does not affect later encoding at all, though it may be used at a later stage of processing. This is consistent with the predictions of raw-cue encoding models that either do not use coarticulation in the frication or use it as evidence for the vowel. This predicts that N1 responses to sounds spoken in different contexts should be the same.

The second possibility is that listeners treat differences in the frication (between /s_L/ and /s_H/) as cues to the fricative and compensate for them when encoding of the vocalic portion. Thus, when spoken in the context of /s_H/, which has higher-frequency spectral information, listeners would treat the onset of the vocalic portion as having a lower frequency that it actually does. Conversely, when spoken with /s_L/, listeners would treat the vocalic onset as having a higher frequency. This is analogous to the context compensation effects predicted for talker and rate differences and could quite clearly be accounted for by auditory contrast (lateral, extrinsic encoding models) via spectral contrast. It could also be accounted for by a model like C-CuRE (feedback, extrinsic encoding) if the frication differences led to differential activation of fricative categories, such as partial activation for /ʃ/ based on the /s_L/ stimulus. In this case, both approaches lead to the same prediction: listeners encode vowel onset cues relative to these coarticulatory differences in the frication. Specifically, these models predict that N1 responses in each context should show the opposite pattern of results to N1 responses to differences in the vocalic portion. Given the results of Experiment 8 showing larger N1s for /i/ than for /u/, we would expect larger N1s for /s_L/ than for /s_H/.

A third possibility is that listeners treat this information as an anticipatory vowel cue and use it to prime potential cue-values. If this is the case, coarticulation in the frication is not treated as contextual variability identifying the fricative, but can still affect encoding of the subsequent vocalic information (via priming rather than compensation for context).

This would be similar to the type of effect seen in audiovisual speech, where visual information can enhance processing of acoustic cues, as evidenced by shorter N1 latencies for audiovisual speech than for audio-only speech (van Wassenhove et al., 2005). This could have one of two effects. First, if listeners use a strategy similar to the relative encoding models proposed for other types of context effects, we would expect the cues in the frication to prime particular vowel categories. In this case, we would expect larger N1s for $/s_H/$ than for $/s_L/$ (the same pattern seen for /i/ [higher frequency] and /u/ [lower frequency]). Second, facilitatory effects could also produce a difference in the N1 as a function of whether the information in the frication and vocalic portions matched. This would be similar to the results seen for audiovisual speech. Specifically, we would expect larger N1s when the two sounds mismatched than when the matched.

To test these three predictions, I ran an experiment in which I examined ERP responses to the vowels /i/ and /u/ in FV syllables. Two fricatives, /s/ and /z/, were used to see whether the effects generalize to multiple coarticulatory contexts.

6.2   Experiment 9: Coarticulatory context effects

In this experiment, listeners heard stimuli consisting of the syllables /si/, /su/, /zi/, and /zu/. In these syllables, coarticulatory effects of vowel rounding produce changes in the frication. This provides listeners with an anticipatory cue about the vowel, allowing us to examine a type of context effect that we haven't looked at in the previous experiments. In addition, we can examine the effect of a standard context effect, the difference between /s/ and /z/ (though the cues to fricative voicing are not likely to lead to strong context effects for /i/ and /u/).

This experiment used a 2AFC task in which each response choice is equi-probable rather than the target detection task used in Experiments 6 and 8. Since the effects of interest relate to the N1 component, a 2AFC task was chosen as it was simpler and we were not concerned with categorization. In addition, Adjar was also used here to help remove

overlap between ERP components to the frication and those to the vocalic portion.

For each fricative-vowel combination, stimuli were cross-spliced to create conditions in which the coarticulatory information in the fricative is either the same or different from the vowel. This allows us to examine several conditions that may affect N1 amplitude. First, we can examine the effect of vocalic differences (/i/ vs. /u/) to see which vowel produces a larger N1. This can then be compared to other effects to look for evidence of context compensation (i.e., if listeners attribute differences in the frication to the fricative) or facilitation (if they attribute them to the vowel).

Second, we can examine the effect of the particular fricative (i.e., $/s_H/$ vs. $/s_L/$ or $/z_H/$ vs. $/z_L/$). This allows us to look for evidence of anticipatory effects due to coarticulation from the vowel. If listeners use this information as a cue to the category of the upcoming vowel, we would expect the coarticulatory context to either reinforce the N1 differences observed for the different vocalic portions via priming (i.e., larger N1s in the context of an /i/). In contrast, if they treat differences in the frication as a source of contextual variability (i.e., a cue to the fricative) and compensate for this effect at the level of cue encoding, we would expect their N1 responses to reflect such compensation (larger N1s in the context of an /u/). Finally, if coarticulatory information has no effect on vocalic encoding, regardless of how the listener uses that information, we would expect N1 responses to be the same across coarticulatory contexts.

Third, we can examine the effect of whether the coarticulatory information in the fricative matched or mis-matched the vowel (the interaction of the two main effects above). This would not reveal the effects of relative encoding based on coarticulatory context (since they would cancel each other out). Instead, it allows us to see whether listeners detect the match between the frication and vocalic portions, independent of the acoustic characteristics of the stimulus, at this early stage of processing. This would be indicative of facilitation of the N1 where listeners used the information in the frication as a cue to the upcoming

vowel. In addition, this may be informative for further establishing the N1 as an index of cue-level encoding.

6.2.1 Methods

6.2.1.1 *Participants*

Twenty-four people participated in the experiment. Participant recruitment, consent, participation criteria, and compensation procedures were the same as the other ERP experiments.

6.2.1.2 *Design*

Participants performed a 2AFC vowel identification task. Each combination of fricative (/s/ or /z/), coarticulation (match or mis-match), and vowel (/i/ or /u/) was presented 45 times, for a total of 360 trials. The experiment took approximately 60 minutes and was completed in a single session.

6.2.1.3 *Stimuli*

The male talker recorded for the talker variability experiments recorded the stimuli for this experiment as well. Several tokens of each of the four syllables (/si/, /su/, /zi/, and /zu/) were recorded. Recordings were made with the same equipment and in the same sound-attenuated room as in previous experiments.

The talker was asked to pronounce each word slowly. This was done to obtain tokens of each fricative that were long enough that they could be shortened by 250 ms (in order to use the Adjar procedure) while still allowing listeners to clearly identify the word. Several tokens of each word were recorded, and the highest-quality token was selected for the experiment.

The length of the frication portions were first adjusted so that they were equal. This was done by first measuring the length of each frication portion to determine which was

the shortest. Then, the other three stimuli were shortened by cutting from the end of the sound at zero-crossings closest to the duration of the shortest stimulus. /zi/ had the shortest frication in the set (699 ms), and the /si/, /su/, and /zu/ tokens were cut by 266, 369, and 128 ms, respectively. The same procedure was used to equate the length of the vocalic portions. Again, /zi/ had the shortest length (508 ms), and the /si/, /su/, and /zu/ tokens were cut by 63, 28, and 77 ms, respectively. Next, the stimuli were cross-spliced to create the coarticulatory mismatch conditions between the fricative and vowel. Finally, the pitch-synchronous overlap-add method was used to shorten each fricative in 10 steps over a 250 ms range so that the ISI between the onset of the fricative and the onset of the vowel varied. The ISIs ranged from 450 to 700 ms.

6.2.1.4   *Procedure, EEG recording, and data processing*

The procedure was the same as Experiment 1, except that participants indicated whether the word they heard had an /i/ or /u/ in it. EEG recording and data processing procedures were the same as previous ERP experiments, and the Adjar procedure was used to reduce component overlap.

6.2.2   Results

Listeners' behavioral responses indicated that they identified each of the vowels highly accurately (mean accuracy: 98.9%).

Before analyzing N1 amplitude to vocalic onsets, the Adjar procedure was applied to reduce overlap from ERP components to the frication. Adjar was run on each of the four frication segments ($/s_H/$, $/s_L/$, $/z_H/$, and $/z_L/$) and converged on a stable estimate of the overlap in each case. However, the pre-stimulus baseline after Adjar was not flat in every case, suggesting some overlap from preceding components was still present. Figure 6.1 shows ERP waveforms for each frication segment time-locked to the onset of the vocalic portion. The amount of overlap was reduced for each segment, but the pre-stimulus baseline

was not flat for the two /z/ segments. Because of this, comparisons made with the /z/ stimuli may reflect processing related to the frication rather than the vocalic portion. Thus, to examine the effect of vowel coarticulation (from information in the frication) on encoding of vocalic information, the /z/ stimuli were not included in the analyses.

Figure 6.2 shows grandaverage ERPs waveforms for the average of the frontal channels as a function of the vocalic portion, time-locked to its onset. The /i/ segments appeared to produce larger N1s that the /u/ segments, consistent with the findings for the /i/-/u/ stimuli in Experiment 8.

Figure 6.3 shows ERPs as a function of the coarticulatory information in the frication (i.e., $/s_H/$ vs. $/s_L/$). Here, N1s appear to be larger for $/s_L/$ than for $/s_H/$. This reversed pattern of results suggests a context compensation effect: listeners compensate for higher frequency sounds produced in the context of an /i/ by treating subsequent sounds as lower than they actually are (producing smaller N1s). This indicates that the cues in the frication were processed as contextual information.

Figure 6.4 shows ERPs as a function of whether the information in the frication and vocalic portions matched. That is, if the frication was $/s_L/$, the matching vocalic portion would be /u/ (lower formant frequencies) and mismatching vocalic portion would be /i/ (higher frequencies). This allows us to examine effects of facilitation from information in the frication on the N1 (i.e., a smaller N1 for the match condition). There does not appear to be much of a difference in the N1, though there does appear to be a large effect in the P2.

To assess these observations statistically, mean N1 amplitude was calculated from 90 to 140 ms after the onset of the vocalic portion. Mean N1 amplitude as a function of vowel and frication context are shown in Figure 6.5. A 2 (vowel) x 2 (coarticulatory context) within-subjects ANOVA was run to evaluate differences in N1 amplitude. There was a main effect of coarticulatory context ($F(1,23)=4.41$, $p=0.047$) with $/s_L/$ sounds producing larger N1s to the vowel than $/s_H/$ sounds. This supports the hypothesis that listeners treat the
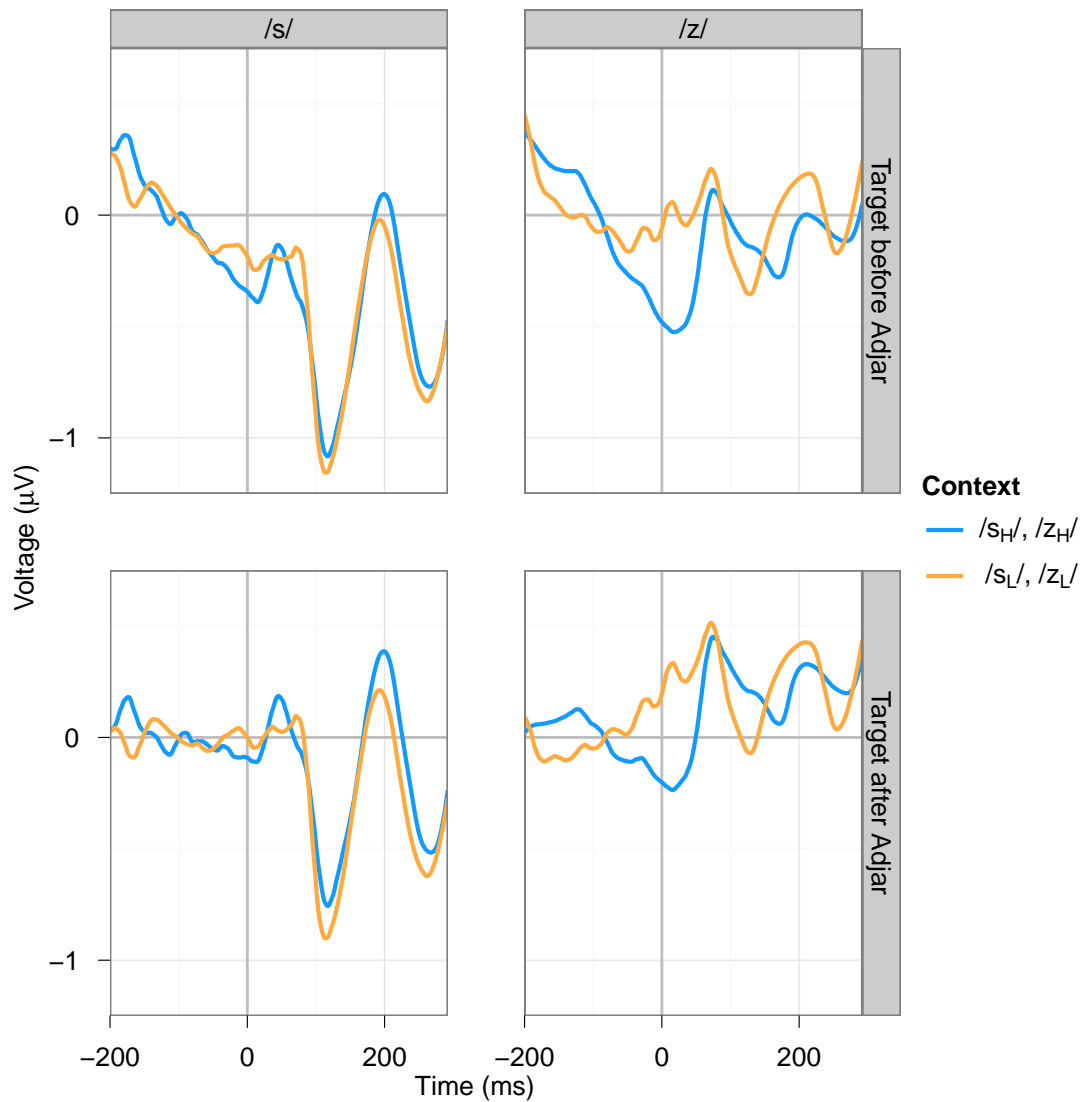
Figure 6.1: Experiment 9 results — effectiveness of Adjar procedure. Grandaverage ERP waveforms time-locked to the onset of the vocalic portion for the average of the three frontal channels before (top panels) and after (bottom panels) the Adjar procedure was applied. Adjar was run separately on each of the four frication segments ($/s_H/$, $/s_L/$, $/z_H/$, and $/z_L/$). For the two /s/ sounds (left panels), the majority of the overlap was removed, but for the /z/ sounds (right panels) there was still considerable overlap from the preceding frication after running Adjar.
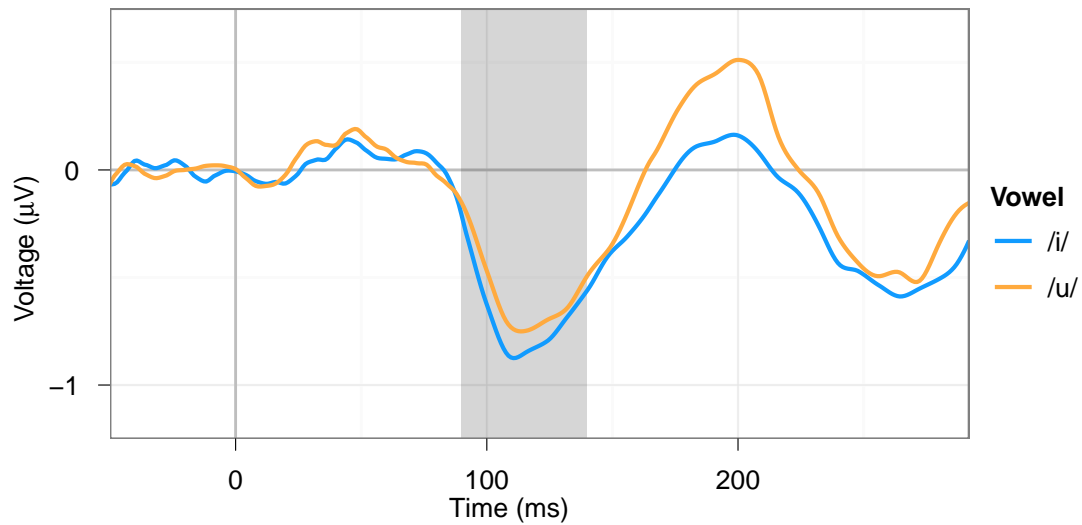
Figure 6.2: Experiment 9 results — ERP waveforms by vocalic sound. Grandaverage ERPs for the average of the three frontal channels to the /i/ and /u/ vocalic stimuli preceded by an /s/. Shaded area indicates time range used to compute mean N1 amplitude.
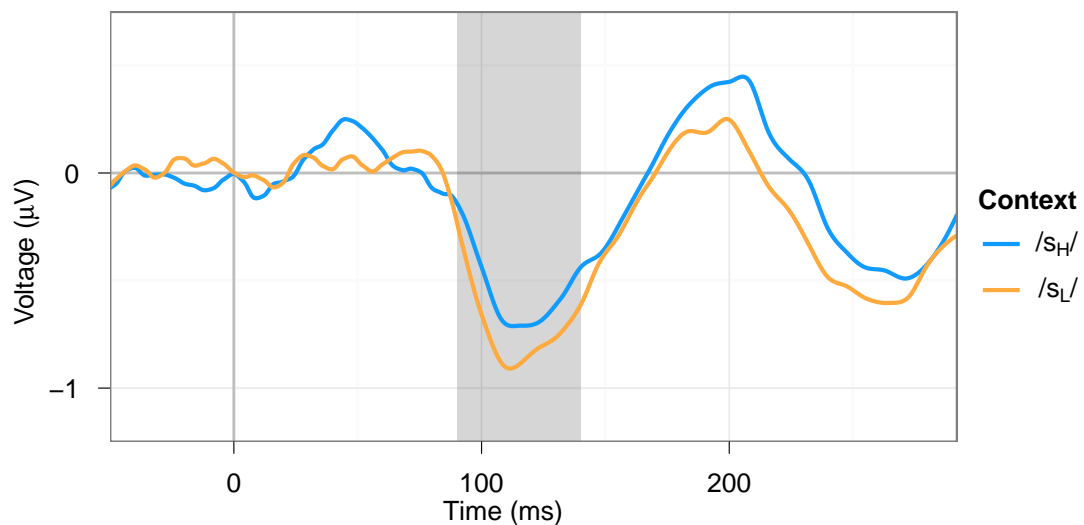
Figure 6.3: Experiment 9 results — ERP waveforms by coarticulatory context. Grandaverage ERPs for the average of the three frontal channels as a function of preceding coarticulatory context ($/s_H/$ vs. $/s_L/$). N1 amplitude was greater for the $/s_L/$ stimuli than for the $/s_H/$ stimuli. Shaded area indicates time range used for mean N1.
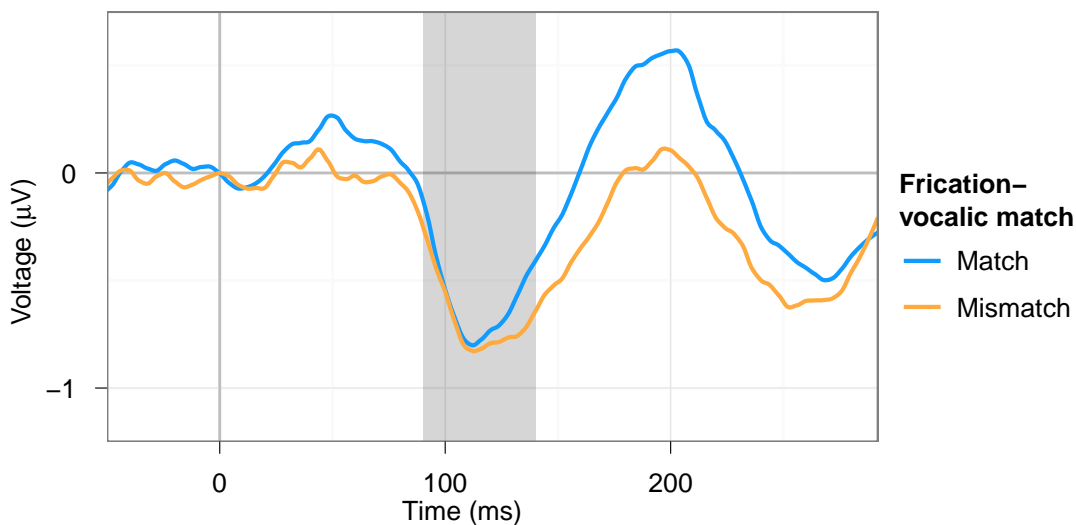
Figure 6.4: Experiment 9 results — ERP waveforms by FV match. Grandaverage ERPs for the average of the frontal channels as a function of whether the vowel information in the frication matched or mismatched the vowel information in the vocalic portion (i.e., $/s_L u/$ and $/s_H i/$ [match] vs. $/s_L i/$ or $/s_H u/$ [mismatch]). Shaded area indicates time range used for mean N1.
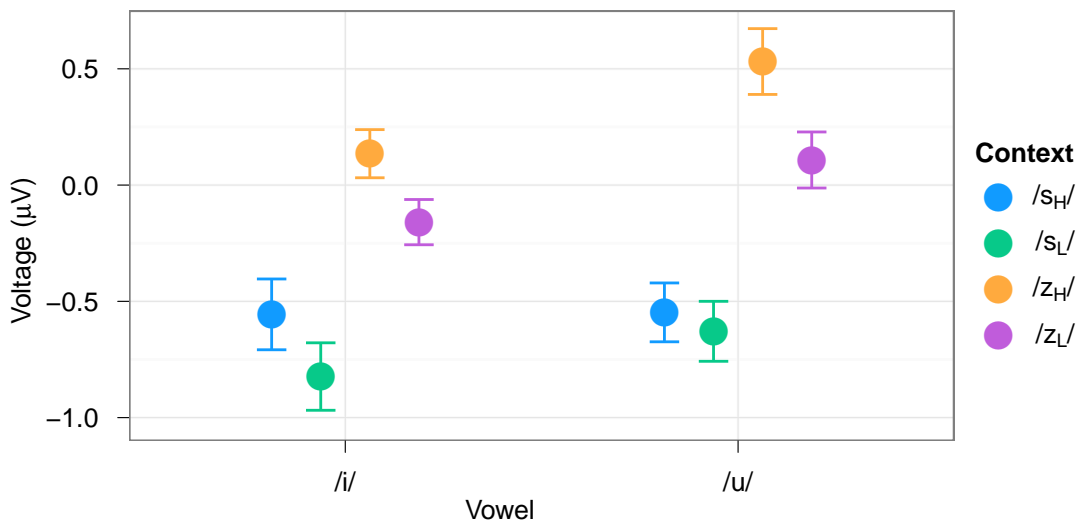
Figure 6.5: Experiment 9 results — N1 amplitude. Mean amplitude as a function of vocalic portion (/i/ vs. /u/) and preceding context ($/s_H/$, $/s_L/$, $/z_H/$, and $/z_L/$). Note that N1 amplitude for the /z/ stimuli should be interpreted with caution, since the Adjar procedure was not effective at reducing component overlap for those stimuli.

coarticulatory information as cues to fricative context, compensating for these differences when processing the vocalic portion. In addition, it suggests that this effect occurs during cue encoding, supporting a relative encoding model.

The effect of vowel was not significant ($F(1,23)=1.14$, $p=0.299$). This suggests that the overall vowel difference did not produce a change in N1 amplitude and appears to contradict the results of previous experiments, particularly Experiment 8, which examined the same vowel distinction and found large effects. There may have been insufficient power to detect a vowel effect with only the /s/ stimuli. Indeed, an analysis including the /z/ stimuli as well, does show a significant effect of vowel ($F(1,23)=8.61$, $p=0.007$). However, this result should be interpreted with the caveat that the /z/ stimuli likely also contained overlap from the fricative (though this comparison averages across the two fricatives, so any overlap should be the same in both conditions).

The interaction was also non-significant ($F(1,23)=1.15$, $p=0.295$). This suggests that N1 amplitude was not sensitive to the overall effect of whether the frication and vocalic portions matched or mismatched. That is, in this case, the differences in the N1 we observed reflect how cues are encoded, not higher level processes that reflect categorization. This also argues against the idea that coarticulatory information is being used to facilitate later cue encoding for these stimuli.

6.2.3   Discussion

The results of this experiment suggest that listeners treated the coarticulatory information as a source of contextual variation, that is, as a cue to the identity of the fricative. This indicates that both sources of anticipatory information in earlier segments (i.e., information due to the intended fricative and anticipatory information due to the upcoming vowel) are used similarly by listeners. In addition, this context information was factored out of listeners subsequent cue estimates, supporting relative encoding models. Thus, it does not appear that listeners treat context effects and additional phonetic cues differently,

though the coarticulatory information about the vowel could have also been used for vowel categorization via feedforward mechanisms.

Fricative context (i.e., /s/ vs. /z/) also provided a source of context information in this experiment, though overlap in ERP components precluded an extensive analysis of the /z/ stimuli. This difference could similarly lead to context compensation. Fricative voicing contrasts vary in F0 and F1 transitions (Stevens et al., 1992), which could provide listeners with coarticulatory information that can be factored out of their vowel estimates. Specifically, the /z/ syllables should have lower F0 and F1 values at the onset of the vocalic portion than the /s/ syllables, and /i/ should have lower slightly F0 and F1 values than /u/. Thus, if listeners compensate for the effect of fricative voicing during cue encoding, we would expect more /u/-like (higher F0 and F1) responses in the context of a /z/ and more /i/-like responses (lower F0 and F1) in the context of an /s/. Indeed, this is exactly the pattern of results that was observed, providing additional support for the relative encoding hypothesis (see mean amplitudes for /z/ in Figure 6.5). However, the large amount of overlap from ERPs to the preceding frication in the /z/ stimuli prevents any strong conclusions from being made here.

Overall, these results fit with the predictions of relative encoding models, particularly auditory contrast approaches (Lotto & Kluender, 1998; Diehl et al., 2004; Kluender, 2003; Holt, 2005). At a very basic level, when the spectral mean of a fricative is low, subsequent formant frequencies appear to be treated as higher (as predicted by Lotto & Kluender, 1998). However, considering these results along with those of the previous experiments, a model that uses feedback to compute relative cues, like C-CuRE (McMurray & Jongman, 2011; Cole et al., 2010), may be more consistent with the overall pattern of results for talker, rate, and coarticulatory context effects. In this sense, the reduced spectral mean of the /u/-conditioned fricatives may have led listeners to partially mis-categorize stimuli as /ʃ/, resulting in differences in how the vowel was interpreted.

The final chapter evaluates these and the other proposed models in light of the results of each set of experiments and discusses what the results suggest about how listeners handle contextual variability during speech perception.

CHAPTER 7
GENERAL DISCUSSION

The search for invariance has been central to research in speech perception (Perkell & Klatt, 1986), and debate about whether listeners have access to context-invariant information continues. A growing consensus is that there is unlikely to be a sufficient set of simple, context-invariant cues that serve as the basis of speech perception. More complex processes for compensating for contextual variability are required. The primary purpose of this dissertation was to evaluate proposed accounts of how listeners handle contextual variability in speech using a novel ERP approach for measuring cue encoding.

Here, I will briefly review the key results of the experiments as well as the predictions made by the general classes of models, discussing how a complete model of speech perception might be able to account for context effects. In addition, I will review the results suggesting that the auditory N1 can be used as an index of cue encoding and the P3 as an index of categorization, and discuss future work using these approaches.

7.1 Summary of results

First, the results of these experiments demonstrated that listeners compensate for variability across different types of context. Experiment 6 found that listeners compensate for differences in talker gender when making voicing judgments, and Experiments 5 and 6 found compensation effects for vowel judgments. In addition, Experiment 7 and the 2AFC task in Experiment 8 both found that listeners compensate for differences in speaking rate when making voicing judgments.

The primary empirical question addressed by these experiments was whether particular types of context information (talker, rate, and coarticulation) affect perception at a cue level (indicated by changes in the auditory N1) or at a category level (indicated by a change in the P3). Experiment 6 examined listeners' responses to differences VOT and F1

cues as a function of talker gender, finding differences in the N1 for VOT that were consistent with context compensation, but no differences for F1. Experiment 8 examined listeners' responses to differences in VOT and a collection of vowel cues (F2, F3, B2, and B3) as a function of speaking rate. An overall effect of rate on both continua was found in the N1, but not in the direction predicted for rate compensation for VOT. Finally, Experiment 9 examined N1 responses to vowel sounds in the context of different coarticulatory information, finding a difference in the N1 that indicates context compensation.

A secondary question was whether N1 and P3 amplitude varied linearly with various acoustic cues and phonological distinctions. Experiments 4 and 6 found some evidence for linear effects of F1 and VOT on the N1, though the effects for F1 were only seen in Experiment 4 and the effect of VOT was marginal when the data were grouped by response. Both experiments found strong evidence for effects on the P3 (relative to listeners' category boundaries). Experiment 8 found additional evidence for VOT effects on the N1 (again, when not grouped by response) and effects of a set of formant cues (F2, F3, B2, and B3) on the N1. This experiment also found strong evidence for effects on the P3 for both types of acoustic cues.

7.2 Evaluating proposed models

Given these results, we can evaluate the classes of models that have been proposed for handling contextual variability. Recall that the two main dimensions that these models differ along are (1) the type of encoding they propose (intrinsic, extrinsic, and raw-cue encoding), and (2) the mechanism by which context information interacts with phonetic cues or categories (feedforward, feedback, or lateral information flow). The experiments presented here allow us to distinguish between models along both of these dimensions.

7.2.1 Intrinsic, extrinsic, and raw-cue encoding

Several of the experiments provide evidence arguing against purely intrinsic approaches. These models suggest that listeners process cues within a phonetic segment in relation to each other (e.g., formant ratios or VOT:VL ratios). However, we saw effects of preceding context, even when these cues were held constant. Experiments 5 and 6 demonstrated this in the case of talker variability for both VOT (as seen in the auditory N1 and in the behavioral responses) and F1 frequency (seen in the behavioral responses). With respect to speaking rate, Experiment 7 and the follow-up to Experiment 8 demonstrated compensation for rate in listeners' behavioral responses. The only case where an argument for purely intrinsic models could be made is in the target detection task of Experiment 8, where SR was varied, but VL was not, and no context effect was observed. However, this seems to be a result of the task, since the same stimuli do show an effect of SR in a 2AFC version. Thus, there is no evidence to support purely intrinsic encoding approaches.

The remaining approaches, extrinsic and raw-cue encoding, suggest two different ways that preceding context information can influence listeners' responses. The distinction between these two approaches lies at the level of cue encoding, where the extrinsic approach predicts that cues are encoded relative to context, while the raw-cue approach predicts they are not. The talker variability and coarticulation experiments addressed this distinction directly, finding support for a relative encoding account, though the results are mixed. Listeners' N1 response to stop consonants varied as a function of whether the preceding carrier phrase was spoken by a man or woman, such that larger N1s (consistent with voiced responses) were seen when the carrier was spoken by a woman. While only a weak effect of VOT was seen in that experiment, clearer effects of VOT were seen in experiments Experiments 4 and 8. Similarly, coarticulatory context had a compensatory effect on listeners encoding of subsequent vowel information, though the overall effect of the vowel was not significant, even though such effects were seen in another experiment (Experiment 8).

In addition, no effects of talker were seen in F1 responses. There are a number of possible reasons for this, including the relatively weak effect of F1 in general and the possibility that effects were obscured by an overall effect of talker gender in the opposite direction. Thus, the results only provide tentative support for a relative encoding process.

Intriguingly, the rate experiments did not provide evidence for relative cue encoding. Those results are more consistent with raw-cue encoding models in which speaking rate has an effect on listeners' phonetic judgments at later stages (e.g., phoneme categorization). Thus, a complete model would need to account for both relative and raw-cue encoding.

7.2.2 Feedforward, feedback, and lateral information flow

Although the coarticulatory results and some of the talker gender results suggest that cues are encoded relative to context information, by themselves, they do not distinguish between lateral and feedback models. However, the speaking rate experiments may lend insight into these issues. First, speaking rate is an extremely likely domain in which we would be able to observe effects of lateral interactions (e.g., via durational contrast). However, we found no compensatory effect of rate on listeners' VOT encoding. This leaves us with feedback models. Feedback from more-abstract, but continuous estimates of rate could have also had an effect on VOT encoding, but again, the results do not support this.

Therefore, this suggests models that use feedback from categories. These models do not predict an effect of rate on cue encoding, which is consistent with the observed results. Because feedback in these models requires discrete categories and rate does not provide listeners with such information (in contrast to talker gender and coarticulatory context), listeners would not have had the information necessary to encode cues relative to speaking rate. Indeed, unsupervised clustering algorithms used to learn phonetic categories (McMurray, Tanenhaus, & Aslin, 2009) would not predict that listeners have multiple categories for a unimodal or uniformly-distributed dimension, and there is evidence that listeners use this type of learning to acquire categories (Maye, Werker, & Gerken, 2002).

More powerful supervised learning algorithms may be able to carve such dimensions into arbitrary categories, but it is not clear what those arbitrary categories correspond to for rate. Since the rate variability experiment found no evidence of relative cue encoding, this supports a model that uses feedback from categories.

As noted above, however, an effect of SR on listeners' behavioral responses was observed. Thus, relative cue encoding cannot be the sole mechanism by which listeners compensate for contextual variability. This is consistent with feedforward models, where SR could be integrated with other acoustic cues directly at the level of phonological features. (This possibility is discussed in more detail in the next section.) Thus, as with the evidence for the type of encoding, the results of these experiments suggest multiple pathways that context information can take during speech perception.

7.3 Building a complete model of speech perception

Together, the results of these experiments support two different approaches to the problem of lack of invariance. Of the existing models, C-CuRE seems to provide the closest explanation for all of the effects observed. It can account for relative cue encoding in the context of different talkers and coarticulatory contexts via feedback. It can also account for rate effects via feedforward mechanisms that treat context information as additional phonetic cues. However, the rate effects seen here are not due to differences in vowel length within a word (which seems like a plausible candidate for a weak phonetic cue to voicing Toscano & McMurray, 2011a), and it seems odd to conclude that sentences spoken quickly cause listeners to simply activate voiceless phonological categories more.

Given this, it may be that listeners do not use an overall estimate of running speaking rate to make phonetic decisions. Instead, they may use durational cues that occur close in time to the phonetic segment (i.e., the initial stop consonant in the target words). This fits with the idea that phonetic segments overlap in time (Fowler, 1984), and it provides an explanation for why vowel length effects would be so much larger than the effects of

running speaking rate. Indeed, Summerfield (1981) and Repp and Lin (1991) both found that preceding rate differences that are temporally adjacent to the stop consonant produce larger effects. Thus, speaking rate may not be a context effect at all. Rather, it may simply be a proxy for the set of durational differences that occur in both the preceding and following vowels that listeners use to distinguish voicing categories (Toscano & McMurray, 2011a). Evaluating this idea will require careful phonetic measurement work as well as further empirical work. However, this seems like a promising way of maintaining the simplicity of feedforward models given that lateral and feedback models do not seem to describe the observed speaking rate effects.

This leads to a model that combines aspects of relative and raw-cue encoding models. Figure 7.1 depicts a possible model that could account for all of these effects. Feedback between levels of processing would allow the model to account for the talker and coarticulation results, while rate effects are handled by bottom-up feedforward activation from a number of cues. This model can handle the results described in the present set of experiments, though it cannot account for all data on context effects in speech perception, such as the influence of nonspeech sounds (Lotto & Kluender, 1998; Holt, 2005).

This type of model could be implemented in a recurrent neural network. For example, normalized recurrence networks (NRN ; Spivey, 2007), can be used to capture the mapping between continuous cue encoding and phoneme categories (McMurray & Spivey, 1999) as well as the effects of multiple cues (or cues and context effects ; Toscano & McMurray, 2008; Clayards & Toscano, 2010). These networks use feedback from category- to cue-level representations to allow activations in the network to settle. Thus, this feedback provides a mechanism for encoding cues relative to expectations from category-level processes. Other models with similar architectures, such as iterative competitive learning (ICL) networks (Mozer, 1990) are functionally similar and could also be used to compute relative cues from category-level information. This relative encoding, along with feedforward
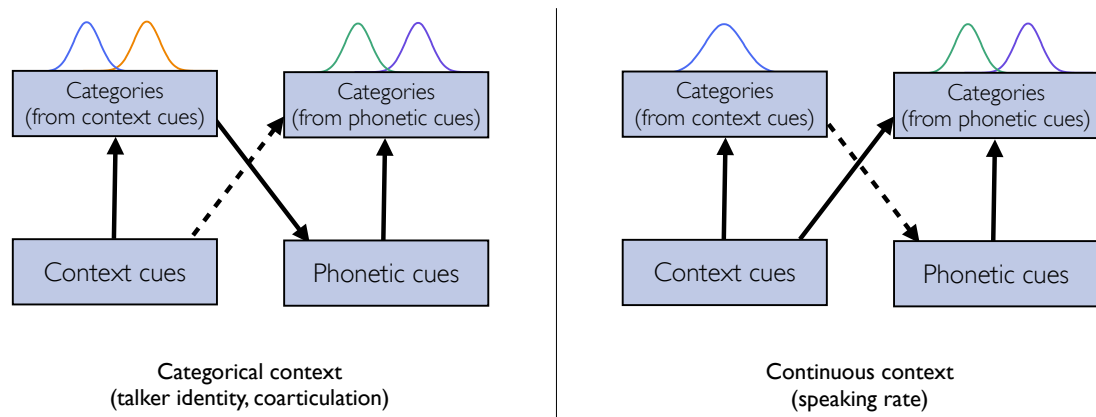
Figure 7.1: Schematic of proposed model of speech perception in context. For context information that provides categorical distinctions (e.g., talker identity, coarticulatory context), information from context categories feeds back to affect encoding of phonetic cues, which is highlighted in the left panel. For context information that provides more continuous estimates, phonetic categories are directly biased, indicated in the right panel.

processes that combine continuous context and cue estimates, suggests that these types of models provide a basic architecture that could be used to implement the general information processing framework outlined above.

7.4 Indexing functions for ERP components

A secondary goal of this dissertation was to provide additional evidence that the auditory N1 can be used as an index of acoustic cue encoding and that the P3 serves as an index of categorization. This distinction is similar to the one made by Picton and Hillyard (1974) for nonspeech sounds and by Toscano et al. (2010) for speech sounds. The experiments in this study extend previous findings by looking at effects of additional acoustic cue dimensions and phonological distinctions. Here, I discuss what these results suggest about the information provided by the N1 and P3.

7.4.1 The auditory N1 as an index of cue encoding and work

7.4.1.1 Cue encoding

Although it is well-documented that the N1 varies with a variety of auditory stimulus properties (Näätanen, 1987), there are also reports that it reflects discontinuities in the processing of speech sounds (Sharma & Dorman, 1999). Moreover, although it was originally proposed that discontinuities in the N1 response to different VOTs may provide a neural basis for differentiating voiced and voiceless sounds (Steinschneider, Schroeder, Arezzo, & Vaughanjr, 1994), more recent evidence suggests that these discontinuities do not map reliably onto listeners' perceptual categories (Sharma et al., 2000) and other researchers have found equal-sized changes in the M100 (N1m) across a VOT continuum (Frye et al., 2007). Further, Toscano et al. (2010) found no evidence of discontinuities along VOT continua with low-amplitude bursts, nor any influence of listeners' phonological categories. This suggests that variation in N1 amplitude reflects continuous acoustic differences in the stimulus.

The results of the current set of experiments provide additional evidence that this is the case for VOT (Experiments 2, 4, and 8), and they show that this applies to naturally-produced speech in addition to synthetic speech. Differences along other acoustic cue dimensions, namely formant frequencies and bandwidths (Experiments 1, 4, 6, 8, and 9), suggest that the N1 also measures encoding of these cues, though the results for F1 were not as robust as those for other cues. Variation between the vowels /ɛ/ and /æ/ produced changes in N1 amplitude in Experiment 1, and variation in F1 along an /ɛ/-/æ/ continuum found a significant linear trend for the target-response trials in Experiments 4. The overall effect was smaller than that for VOT, which is likely due, at least in part, to the small amount of overall variation in N1 amplitude for these stimuli. However, there was no evidence though that variation in N1 amplitude instead reflected categorical differences along that dimension. Linear effects were seen for changes along the F2/F3/B2/B3 continuum from /i/ to /u/ in Experiment 8, suggesting that the N1 reflects encoding of differences in formants more

generally. In addition, there was no evidence that these effects were the result of averaging across categorical responses.

Interestingly, the overall direction of the effects did not correspond to the mean frequency in the sounds. This is seen most clearly in Experiment 8, where the largest N1s were produced by short (i.e., lower frequency) VOTs and high F2 and F3 frequencies. Thus, differences in N1 amplitude do not seem to simply reflect overall changes in acoustic frequency, providing further evidence that the N1 may track differences along specific acoustic cue dimensions. This suggests that variation in the N1 as a function of differences along these continua may correspond to different neural populations that code for sounds in different frequency ranges (Lauter et al., 1985; Bertrand et al., 1991; Picton et al., 1978). This suggests that the basis of the acoustic cues in speech may fundamentally correspond to differences in specific frequency bands. Indeed, many successful approaches to speech coding in automatic speech recognition and communication systems are based on spectral differences in particular frequency bands (e.g., Mel-frequency cepstral coefficients; S. Davis & Mermelstein, 1980).

7.4.1.2   Amount of work

In addition to these effects, it is clear that the N1 indexes more than acoustic differences in the stimulus. This is demonstrated by effects of attention on the N1 (Hansen & Hillyard, 1980; Öhman & Lader, 1972; Picton & Hillyard, 1974), as well as effects of visual information in audiovisual speech perception (Pilling, 2009; van Wassenhove et al., 2005; Besle et al., 2004). Decreases in the amplitude and latency of the N1 for audiovisual speech, for example, may reflect facilitation of speech processing given additional information (van Wassenhove et al., 2005). In addition, the present set of experiments suggests that preceding context has an effect on N1 amplitude and that this effect is due to feedback from higher-level stages of processing. These results also indicate that the N1 reflects processes that are relevant for speech perception.

Recently, many researchers have begun to characterize later ERP components (e.g., the P3 and N4 [N400]) as indices of the amount of work the system is doing for a particular cognitive process. For example, the N400 provides a measure of the amount of work being done to determine the semantic properties of a visual or auditory stimulus. Thus, the N400 is larger for stimuli that violate the subject's semantic expectation (e.g., when reading a sentence like "Look both ways before crossing the *hat*"; Kutas & Hillyard, 1980).

We may be able to characterize the auditory N1 in a similar way. Indeed, the attentional and audiovisual speech effects described above fit with this description. In addition, some of the effects of contextual information fit as well. For example, larger N1s were observed in the context of fast speech, regardless of the particular stimulus, suggesting that listeners may be doing more work to encode acoustic information when a talker speaks quickly. Thus, context can potentially have multiple effects on the N1: in some cases leading to compensation for contextual variability when encoding particular stimuli, and in other cases leading to differences in the overall amount of work done to encode the stimulus.

Although the amount of work being done provides an explanation for some effects on the N1, it seems unlikely that it can account for the effects of the acoustic properties of the stimulus as well. Such an explanation would suggest that listeners do more work to encode voiced sounds (shorter VOTs), as well as back and closed vowels (higher formant frequencies). However, listeners are equally good at categorizing these stimuli and stimuli that produce smaller N1s, and there is no evidence that acoustic information in these frequency ranges is more or less difficult to encode.

7.4.1.3 Summary

As the discussion above illustrates, the data suggest that the N1 reflects both the acoustic properties of the stimulus and the influence of other factors related to auditory processing. Thus, the N1 is best described as an index of both (1) the representation of incoming acoustic information, and (2) the amount of work done to encode that information,

reflecting effects of attention and information from other stages of processing. This fits with the results of previous studies (Picton & Hillyard, 1974; Toscano et al., 2010) as well as the present experiments.

7.4.2   The P3 as an index of categorization

Several of the experiments presented here also measured the P3 component in response to different speech sounds (Experiments 4, 6, and 8), demonstrating that a P3 is consistently produced in response to speech sounds corresponding to a target word with a low probability of occurrence and that the amplitude of the P3 decreases with the acoustic distance from that target word. This extends the results of Toscano et al. (2010), which showed an effect of the P3 for /b/-/p/ and /d/-/t/ distinctions, finding effects for /ɛ/-/æ/ and /i/-/u/ continua as well.

These results fit with the proposal that the P3 serves as an index of categorization or decision-related processes (Picton & Hillyard, 1974; Toscano et al., 2010). P3 amplitude is not directly related to the acoustic properties of the stimulus, but rather is determined by parameters of the task, which are ultimately based on the phonological and lexical categories of interest in the experiment. Differences in the P3 as a function of how well the acoustic stimulus supports the phonological category of the target further suggest that the P3 responds to differences in phonological information, since phonological categories are well-known to be similarly graded (Andruksi, Blumstein, & Burton, 1994; Miller, 1997; Miller & Volaitis, 1989; McMurray et al., 2002). This suggests it can be used as a measure of phonological categorization, similar to discrimination measures (Pisoni & Lazarus, 1974) and eye-movements (McMurray et al., 2002), which also reflect acoustic differences within a phonological category. Overall, these results suggest that the P3 reflects task-defined decision processes, which, in the present experiments, corresponds to phonological categorization.

This is made more apparent by taking differences in individual subjects' category boundaries and responses into account. In addition, since listeners sometimes produced

"target" responses for stimuli that corresponded to the opposite phonetic category for them (i.e., responding "target" for a stimulus with a VOT of 10 ms [corresponding to /b/] when *peach* was the target), we were able to look at differences in P3 amplitude across entire acoustic continua when stimuli were all classified as belonging to a single phonological category. A linear effect of the acoustic cue dimension was found in every case (VOT, F1, and F2/F3/B2/B3 differences). This indicates that listeners are sensitive to acoustic differences within a category, even for sounds that cross their category boundary. This argues strongly against models of categorical perception and fits with the view that speech perception is better characterized as continuous (Carney, Widin, & Viemeister, 1977; Pisoni & Lazarus, 1974; McMurray, Aslin, et al., 2008; Miller, 1997; Andruksi et al., 1994; Massaro & Cohen, 1983; McMurray et al., 2002; Toscano et al., 2010; Gerrits & Schouten, 2004).

Interestingly, strong evidence of context compensation was not seen in P3 responses, even when it was observed in the N1 and in listeners' overt responses. An interaction between talker gender context and distance from the target endpoint along the F1 continuum was observed in Experiment 6, suggesting that P3 amplitude may be greater for stimuli close to the target endpoint when the context leads to responses consistent with that endpoint. However, similar effects were not observed for the other acoustic cue dimensions.

This suggests that the P3 may not completely reflect listeners' categorization of the stimulus, since they clearly take context information into account when recognizing speech. An alternative possibility is that we may have been unable to observe these relatively small context effects in the P3. Indeed, for the stimulus combinations where context effects are expected to be largest (i.e., near category boundaries), our P3 measure is the least sensitive. This suggests that the P3 may serve as a complement to other online measures of categorization, like eye-movements (Tanenhaus et al., 1995; McMurray et al., 2002) and identification tasks, that are more sensitive to differences near category boundaries and do show context effects (McMurray, Clayards, et al., 2008; Toscano & McMurray, 2011a).

Overall, the data suggest that the P3 is closely related to differences in listeners' phonological categories. Thus, the P3 serves as an index of categorization and reflects variation within a category. However, it may not show (or it may be difficult to observe) effects of all factors that influence listeners' behavioral responses.

7.5 Future work

In addition to the effects of talker, rate, and coarticulatory variability that were examined here, other sources of preceding information in spoken language processing can be used by listeners to predict upcoming acoustic information. This includes prosodic cues (Cole, Kim, Choi, & Hasegawa-Johnson, 2007) as well as semantic information (Bicknell, Elman, Hare, McRae, & Kutas, 2010). Expectations from semantic context can also affect speech perception (Connine, 1987). Do these effects occur at the level of cue encoding? Feedback from abstract semantic representations would be a potential source of information for this, though here, information from feedback could lead to priming effects rather than the compensatory effects observed for contextual differences.

Feedback has been a hotly debated topic in work on spoken word recognition as well. Completely feedforward models of word recognition (e.g., Merge; Norris, 2000) allow for veridical perception and prevent "hallucinations" from top-down information on the input signal. Models with feedback (TRACE; McClelland & Elman, 1986), however, allow preceding information to constrain interpretation of the input and improve recognition. The key difference between these types of models lies at the level of either cue encoding or intermediate stages of processing. In a feedforward system, lower-level representations would not be influenced by preceding information, though this information may still have an effect on lexical processing. In a system with feedback, lower-level representations themselves could be affected by lexical information.

Feedback to cue-level representations represents the strongest test of these two approaches. This could be addressed using semantic priming. Listeners would be presented

with a visual prime (a picture of an object) followed by a semantically-related auditory stimulus varying along a single acoustic cue dimension. For example, if the visual prime is a picture of a *peach*, the auditory stimulus would be drawn from a *beach-peach* continuum. The subject would then be presented with an instruction to respond whether either the visual or auditory stimulus started with a /b/ or /p/. If N1 amplitude varies as a function of the preceding visual prime, it would suggest that feedback has an effect on encoding. Specifically, if feedback is used to constrain the interpretation of the incoming signal, we would expect N1 amplitude to show more /b/-like responses (i.e., larger amplitude) when preceded by a /b/ prime, and more /p/-like responses (smaller amplitude) when preceded by a /p/ prime. Future experiments are planned to study these effects.

7.6   Conclusions

Together, the results of this study demonstrate that listeners' compensate for contextual differences in talker gender, speaking rate, and coarticulation. We found preliminary evidence suggesting that acoustic cues are encoded relative to context via feedback from categories (e.g., differences in talker gender), even if those differences are incorrectly attributed to contextual variability (as in the case of coarticulatory context). Moreover, listeners seem to treat continuous sources of context information (like speaking rate) differently, suggesting that context effects observed in these situations may ultimately reflect listeners use of multiple acoustic cues via feedforward activation.

In addition, we found new evidence for continuous cue encoding and graded categorization, building on a growing consensus that speech perception is fundamentally continuous and that listeners do not discard acoustic information in the process of recognizing speech. This also allowed us to further develop the ERP techniques used to assess cue encoding and phonological categorization. Overall, these findings point towards models that are highly sensitive to fine-grained acoustic information and factor out contextual variability during cue encoding using feedback from category-level information.

REFERENCES

Ainsworth, W. (1975). Intrinsic and extrinsic factors in vowel judgments. In *Auditory Analysis and Perception of Speech* (pp. 103–113). London: Academic Press.

Alfonso, P. J., & Baer, T. (1982). Dynamics of vowel articulation. *Language and Speech*, *25*, 151–173.

Allen, J. S., & Miller, J. L. (1999). Effects of syllable-initial voicing and speaking rate on the temporal characteristics of monosyllabic words. *Journal of the Acoustical Society of America*, *106*, 2031–2039.

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, *38*, 419–439.

Andersen, N. (1974). On the calculation of filter coefficients for maximum entropy spectral analysis. *Geophysics*, *39*, 69–72.

Andruksi, J. E., Blumstein, S. E., & Burton, M. (1994). The effect of subphonetic differences on lexical access. *Cognition*, *52*, 163–187.

Antinoro, F., & Skinner, P. H. (1968). The effects of frequency on the auditory evoked response. *Journal of Auditory Research*, *8*, 119–123.

Bates, D., & Sarkar, D. (2011). lme4: Linear mixed-effects models using S4 classes [Computer software manual].

Beckman, J., Helgason, P., McMurray, B., & Ringen, C. (2011). Rate effects on Swedish VOT: Evidence for phonological overspecification. *Journal of Phonetics*, *39*, 39–49.

Bertrand, O., Perrin, F., & Pernier, J. (1991). Evidence for a tonotopic organization of the auditory cortex observed with auditory evoked potentials. *Acta Otolaryngologica*, *491*, 116–122.

Besle, J., Fort, A., Delpuech, C., & Giard, M. (2004). Bimodal speech: early suppressive visual effects in human auditory cortex. *European Journal of Neuroscience*, *20*, 2225–2234.

Bicknell, K., Elman, J. L., Hare, M., McRae, K., & Kutas, M. (2010). Effects of event knowledge in processing verbal arguments. *Journal of Memory and Language*, *63*, 489–505.

Blumstein, S. E., & Stevens, K. N. (1979). Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *Journal of the Acoustical Society of America*, *66*, 1001–1017.

Boersma, P., & Weenink, D. (2010). *Praat: Doing phonetics by computer.* Available from http://www.praat.org/

Boucher, V. J. (2002). Timing relations in speech and the identification of voice-onset times: A stable perceptual boundary for voicing categories across speaking rates. *Perception and Psychophysics*, *64*, 121–130.

Bradlow, A. R., Toretta, G. M., & Pisoni, D. B. (1995). *Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics* (Research on Spoken Language Processing Progress Report No. 20). Bloomington, IN: Indiana University.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*, 433–436.

Butler, R. A. (1973). The cumulative effects of different stimulus repitition rates on the auditory evoked response in man. *Electroencephalography and Clinical Neurophysiology*, *35*, 337–345.

Carney, A. E., Widin, G. P., & Viemeister, N. F. (1977). Noncategorical perception of stop consonants differing in VOT. *Journal of the Acoustical Society of America*, *62*, 961–970.

Cho, T. (2004). Prosodically conditioned strengthening and vowel-to-vowel coarticulation in English. *Journal of Phonetics*, *32*, 141–176.

Cho, T., & Ladefoged, P. (1999). Variation and universals in VOT: evidence from 18 languages. *Journal of Phonetics*, *27*, 207–229.

Christovich, L. A., & Lublinskaya, V. V. (1979). The 'center of gravity' effect in vowel spectra and critical distance between the formants: Psychoacoustical study of the perception of vowel-like stimuli. *Hearing Research*, *1*, 185–195.

Clayards, M., & Toscano, J. C. (2010, August). Modeling age of exposure in L2 learning of vowel categories. Poster presented at the 32nd Annual Conference of the Cognitive Science Society, Portland, OR.

Clopper, C. G., & Pisoni, D. B. (2004). Some acoustic cues for the perceptual categorization of American English regional dialects. *Journal of Phonetics*, *32*, 111–140.

Clopper, C. G., Pisoni, D. B., & de Jong, K. (2005). Acoustic characteristics of the vowel systems of six regional varieties of American English. *Journal of the Acoustical Society of America*, *118*, 1661–1676.

Cole, J., Kim, H., Choi, H., & Hasegawa-Johnson, M. (2007). Prosodic effects on acoustic cues to stop voicing and place of articulation: Evidence from Radio News speech. *Journal of Phonetics*, *35*, 180–209.

Cole, J., Linebaugh, G., Munson, C. M., & McMurray, B. (2010). Unmasking the acoustic effects of vowel-to-vowel coarticulation: A statistical modeling approach. *Journal of Phonetics*, *38*, 167–184.

Connine, C. M. (1987). Constraints on interactive processes in auditory word recognition: The role of sentence context. *Journal of Memory and Language*, *26*, 527–538.

Conrey, B., Potts, G. F., & Niedzielski, N. A. (2005). Effects of dialect on merger perception: ERP and behavioral correlates. *Brain and Language*, *95*, 435–449.

Dahan, D., & Gareth Gaskell, M. (2007). The temporal dynamics of ambiguity resolution: Evidence from spoken-word recognition. *Journal of Memory and Language*, *57*, 483–501.

Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001). Time course of frequency effects in spoken-word recognition: evidence from eye movements. *Cognitive Psychology*, *42*, 317–367.

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.

Daniloff, R., & Moll, K. (1974). Coarticulation of lip rounding. In N. J. Lass (Ed.), *Experimental phonetics.* MSS Information Corporation.

Davis, H., & Zerlin, S. (1966). Acoustic relations of the human vertex potential. *Journal of the Acoustical Society of America*, *39*, 109–116.

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Acoustics, Speech and Signal Processing*, *28*, 357–366.

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

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*, 9–21.

Denes, P. (1955). Effect of duration on the perception of voicing. *Journal of the Acoustical Society of America*, *27*, 761–764.

Di Benedetto, M. (1989). Frequency and time variations of the first formant: Properties relevant to the perception of vowel height. *Journal of the Acoustical Society of America*, *86*, 67.

Diehl, R. L., & Kluender, K. R. (1989). On the Objects of Speech Perception. *Ecological Psychology*, *1*, 121–144.

Diehl, R. L., Lotto, A. J., & Holt, L. L. (2004). Speech perception. *Annual Review of Psychology*, *55*, 149–179.

Dupoux, E., & Green, K. (1997). Perceptual adjustment to highly compressed speech: effects of talker and rate changes. *Journal of Experimental Psychology: Human Perception and Performance*, *23*, 914–927.

Fant, G. (1960). *Acoustic Theory of Speech Production.* The Hague: Mouton.

Forrest, K., Weismer, G., Milenkovic, P., & Dougall, R. N. (1988). Statistical analysis of word-initial voiceless obstruents: Preliminary data. *Journal of the Acoustical Society of America*, *84*, 115–123.

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

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

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

Frye, R. E., Fisher, J. M., Coty, A., Zarella, M., Liederman, J., & Halgren, E. (2007). Linear coding of voice onset time. *Journal of Cognitive Neuroscience*, *19*, 1476–1487.

Fujisaki, H., & Kawashima, T. (1968). The roles of pitch and higher formants in the perception of vowels. *IEEE Transactions on Audio and Electroacoustics*, *AU-16*, 73–77.

Fujisaki, H., & Kunisaki, O. (1978). Analysis, recognition, and perception of voiceless fricative consonants in Japanese. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *26*, 21–27.

Gelfer, M. P., & Mikos, V. A. (2005). The relative contributions of speaking fundamental frequency and formant frequencies to gender identification based on isolated vowels. *Journal of Voice*, *19*, 544–54.

Gerrits, E., & Schouten, M. E. H. (2004). Categorical perception depends on the discrimination task. *Perception and Psychophysics*, *66*, 363–376.

Gerstman, L. (1968). Classification of self-normalized vowels. *IEEE Transactions on Audio and Electroacoustics*, *16*, 78–80.

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

Hagiwara, R. (1997). Dialect variation and formant frequency: The American English vowels revisited. *Journal of the Acoustical Society of America*, *102*, 655–658.

Hansen, J. C., & Hillyard, S. A. (1980). Endogenous brain potentials associated with selective auditory attention. *Electroencephalography and Clinical Neurophysiology*, *49*, 277–90.

Heinz, J. M. (1961). On the properties of voiceless fricative consonants. *Journal of the Acoustical Society of America*, *33*, 589.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, *97*, 3099–3111.

Holt, L. L. (2005). Temporally nonadjacent nonlinguistic sounds affect speech categorization. *Psycholgical Science*, *16*, 305–312.

Hopfinger, J. B., & Mangun, G. R. (1998). Reflexive attention modulates processing of visual stimuli in human extrastriate cortex. *Psychological Science*, *9*, 441–447.

House, A. S., & Fairbanks, G. (1953). The influence of consonant environment upon the secondary acoustical characteristics of vowels. *Journal of the Acoustical Society of America*, *25*, 105–113.

Huber, J. E., Stathopoulos, E. T., Curione, G. M., Ash, T. A., & Johnson, K. (1999). Formants of children, women, and men: The effects of vocal intensity variation. *Journal of the Acoustical Society of America*, *106*, 1532–1542.

Hughes, G. W., & Halle, M. (1956). Spectral Properties of Fricative Consonants. *Journal of the Acoustical Society of America*, *28*, 303–310.

Jaeger, T. F. (2008). Categorical data analysis: Away from anovas (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*, 434–446.

Jenkins, J. J., & Strange, W. (1999). Perception of dynamic information for vowels in syllable onsets and offsets. *Perception, and Psychophysics*, *61*, 1200–1210.

Jenkins, J. J., Strange, W., & Miranda, S. (1994). Vowel identification in mixed-speaker silent-center syllables. *Journal of the Acoustical Society of America*, *95*, 1030–1043.

Johnson, K. (1991). Differential effects of speaker and vowel Variability on fricative perception. *Language and Speech*, *34*, 265–279.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson & J. W. Mullenix (Eds.), *Talker Variability in Speech Processing* (pp. 145–165). London: Academic Press.

Johnson, K., Strand, E. A., & D'Imperio, M. (1999). Auditory-visual integration of talker gender in vowel perception. *Journal of Phonetics*, *27*, 359–384.

Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. *Journal of the Acoustical Society of America*, *108*, 1252–63.

Kessinger, R. H., & Blumstein, S. E. (1997). Effects of speaking rate on voice-onset time in Thai, French, and English. *Journal of Phonetics*, *25*, 143–168.

Kessinger, R. H., & Blumstein, S. E. (1998). Effects of speaking rate on voice-onset time and vowel production: Some implications for perception studies. *Journal of Phonetics*, *26*, 117–128.

Kewley-Port, D. (1982). Measurement of formant transitions in naturally produced stop consonant-vowel syllables. *Journal of the Acoustical Society of America*, *72*, 379–89.

Kluender, K. (2003). Sensitivity to change in perception of speech. *Speech Communication*, *41*, 59–69.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207*, 203–205.

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America*, *29*, 98–104.

Lauter, J. L., Herscovitch, P., Formby, C., & Raichle, M. E. (1985). Tonotopic organization in human auditory cortex revealed by positron emission tomography. *Hearing Research*, *20*, 199–205.

Lawson, E. A., & Gaillard, A. W. K. (1981). Evoked potentials to consonant-vowel syllables. *Acta Psychologica*, *49*, 17–25.

Lewis, D. (1936). Vocal resonance. *Journal of the Acoustical Society of America*, *8*, 91–99.

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

Liberman, A. M., Delattre, P., & Cooper, F. S. (1952). The role of selected stimulus-variables on the perception of unvoiced stop consonants. *American Journal of Psychology*, *65*, 497–516.

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

Lindblom, B. (1996). Role of articulation in speech perception: Clues from production. *Journal of the Acoustical Society of America*, *99*, 1683–1692.

Lisker, L. (1986). "Voicing" in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees. *Language and Speech*, *29*, 3–11.

Lisker, L., & Abramson, A. (1964). A cross-linguistic study of voicing in initial stops: Acoustical measurements. *Word*, *20*, 384–422.

Lobanov, B. M. (1971). Classification of russian vowels spoken by different speakers. *Journal of the Acoustical Society of America*, *49*, 606–608.

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

Luck, S. J., & Lopez-Calderon, J. (2010). *ERPLAB Toolbox*. Available from http://www.erpinfo.org/erplab/

Magen, H. S. (1997). The extent of vowel-to-vowel coarticulation in English. *Journal of Phonetics*, *25*, 187–205.

Magnuson, J. S., Dixon, J. A., Tanenhaus, M. K., & Aslin, R. N. (2007). The dynamics of lexical competition during spoken word recognition. *Cognitive Science*, *30*, 133–156.

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

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

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

Martin, B. A., Sigal, A., Kurtzberg, D., & Stapells, D. R. (1997). The effects of decreased audibility produced by high-pass noise masking on cortical event-related potentials to speech sounds/ba/ and /da/. *Journal of the Acoustical Society of America*, *101*, 1585–1599.

Martin, J. G., & Bunnell, H. T. (1981). Perception of anticipatory coarticulation effects. *Journal of the Acoustical Society of America*, *69*, 559–67.

Massaro, D. W., & Cohen, M. M. (1976). The contribution of fundamental frequency and voice onset time to the /zi/-/si/ distinction. *Journal of the Acoustical Society of America*, *60*, 704–717.

Massaro, D. W., & Cohen, M. M. (1983). Categorical or continuous speech perception: A new test. *Speech Communication*, *2*, 15–35.

Massaro, D. W., & Oden, G. C. (1980). Speech perception: A framework for research and theory. In N. Lass (Ed.), *Speech and Language: Advances in Basic Research and Practice* (Vol. 3, pp. 129–165). New York: Academic Press.

May, J. (1976). *Vocal tract normalization for /s/ and /ʃ/* (Status Report on Speech Research No. SR-48). New Haven, CT: Haskins Laboratories.

Maye, J., Werker, J., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*, B101–B111.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*, 1–86.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746–748.

McMurray, B., Aslin, R. N., Tanenhaus, M. K., Spivey, M., & Subik, D. (2008). Gradient sensitivity to within-category variation in speech: Implications for categorical perception. *Journal of Experimental Psychology: Human Perception and Performance*, *34*, 1609–1631.

McMurray, B., Aslin, R. N., & Toscano, J. C. (2009). Statistical learning of phonetic categories: insights from a computational approach. *Developmental Science*, *12*, 369–378.

McMurray, B., Clayards, M. A., Tanenhaus, M. K., & Aslin, R. N. (2008). Tracking the time course of phonetic cue integration during spoken word recognition. *Psychonomic Bulletin and Review*, *15*, 1064–1071.

McMurray, B., Cole, J. S., & Munson, C. M. (in press). Features as an emergent product of computing perceptual cues relative to expectations. In R. Ridouane & N. Clement (Eds.), *Where do features come from?*

McMurray, B., & Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review*, *118*, 219–46.

McMurray, B., Samelson, V. M., Lee, S. H., & Tomblin, J. B. (2010). Individual differences in online spoken word recognition: Implications for SLI. *Cognitive Psychology*, *60*, 1–39.

McMurray, B., & Spivey, M. J. (1999). The categorical perception of consonants: The interaction of learning and processing. In *Proceedings of the Chicago Linguistics Society* (Vol. 35, pp. 205–220).

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition*, *86*, B33–42.

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2009). Within-category VOT affects recovery from "lexical" garden paths: Evidence against phoneme-level inhibition. *Journal of Memory and Language*, *60*, 65–91.

Mermelstein, P. (1978). On the relationship between vowel and consonant identification when cued by the same acoustic information. *Perception and Psychophysics*, *23*, 331–336.

Miller, J. L. (1981). Effects of speaking rate on segmental distinctions. In P. D. Eimas & J. L. Miller (Eds.), *Perspectives on the study of speech* (pp. 39–74). Hillsdale, NJ: Lawrence Erlbaum.

Miller, J. L. (1997). Internal structure of phonetic categories. *Language and Cognitive Processes*, *12*, 865–870.

Miller, J. L., & Dexter, E. R. (1988). Effects of speaking rate and lexical status on phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, *14*, 369–378.

Miller, J. L., Green, K. P., & Reeves, A. (1986). Speaking rate and segments: A look at the relation between speech production and speech perception for the voicing contrast. *Phonetica*, *43*, 106–115.

Miller, J. L., & Grosjean, F. (1981). How the components of speaking rate influence perception of phonetic segments. *Journal of Experimental Psychology: Human Perception and Performance*, *7*, 208–215.

Miller, J. L., Grosjean, F., & Lomanto, C. (1984). Articulation rate and its variability in spontaneous speech: A reanalysis and some implications. *Phonetica*, *41*, 215–225.

Miller, J. L., & Liberman, A. M. (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel. *Perception and Psychophysics*, *25*, 457–65.

Miller, J. L., & Volaitis, L. E. (1989). Effect of speaking rate on the perceptual structure of a phonetic category. *Perception and Psychophysics*, *46*, 505–512.

Miller, J. L., & Wayland, S. C. (1993). Limits on the limitations of context-conditioned effects in the perception of [b] and [w]. *Perception and Psychophysics*, *54*, 205–10.

Mordkoff, J. T., & Grosjean, M. (2001). The lateralized readiness potential and response kinetics in response-time tasks. *Psychophysiology*, *38*, 777–786.

Morris, R. J., McCrea, C. R., & Herring, K. D. (2008). Voice onset time differences between adult males and females: Isolated syllables. *Journal of Phonetics*, *36*, 304–317.

Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, *9*, 453–467.

Mozer, M. C. (1990). Discovering discrete distributed representations with iterative competitive learning. In *Proceedings of the 1990 conference on advances in neural information processing systems 3* (pp. 627–634). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Näätanen, R. (1987). The N1 wave of the human electric and magnetic response to sound: A review and analysis of the component structure. *Psychophysiology*, *24*, 375–425.

Nearey, T. M. (1986). Modeling the role of inherent spectral change in vowel identification. *Journal of the Acoustical Society of America*, *80*, 1297–1308.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America*, *85*, 2088–2113.

Nearey, T. M. (1990). The segment as a unit of speech perception. *Journal of Phonetics*, *18*, 347–373.

Nearey, T. M. (1997). Speech perception as pattern recognition. *Journal of the Acoustical Society of America*, *101*, 3241–3254.

Niedzielski, N. (1999). The effect of social information on the perception of sociolinguistic variables. *Journal of Language and Social Psychology*, *18*, 62–85.

Oden, G. C., & Massaro, D. W. (1978). Integration of featural information in speech perception. *Psychological Review*, *85*, 172–191.

Ohala, J. J. (1996). Speech perception is hearing sounds, not tongues. *Journal of the Acoustical Society of America*, *99*, 1718–1725.

Öhman, A., & Lader, M. (1972). Selective attention and "habituation" of the auditory averaged evoked response in humans. *Physiology and Behavior, 8*, 79–85.

Öhman, S. E. G. (1966). Coarticulation in VCV utterances: Spectrographic measurements. *Journal of the Acoustical Society of America*, *39*, 151–168.

Ostreicher, H. J., & Sharf, D. J. (1967). Effects of coarticulation on the identification of deleted consonant and vowel sounds. *Journal of Phonetics*, *4*, 285–301.

Perkell, J. S., & Klatt, D. H. (Eds.). (1986). *Invariance and Variability in Speech Processes*. Hillsdale, NJ: Lawrence Erlbaum.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, *24*, 175–184.

Picton, T. W., & Hillyard, S. a. (1974). Human auditory evoked potentials. II. Effects of attention. *Electroencephalography and clinical neurophysiology*, *36*, 191–9.

Picton, T. W., Woods, D. L., & Proulx, G. B. (1978). Human auditory sustained potentials. II. Stimulus relationships. *Electroencephalography and Clinical Neurophysiology*, 198–210.

Pilling, M. (2009). Auditory event-related potentials (ERPs) in audiovisual speech perception. *Journal Speech, Language, and Hearing Research*, *52*, 1073–1081.

Pind, J. (1995). Speaking rate, voice-onset time, and quantity: The search for higher-order invariants for two Icelandic speech cues. *Perception and psychophysics*, *57*, 291–304.

Pisoni, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception and Psychophysics*, *13*, 253–260.

Pisoni, D. B. (1993). Long-term memory in speech perception: Some new findings on talker variability, speaking rate and perceptual learning. *Speech Communication*, *13*, 109–125.

Pisoni, D. B., Carrell, T. D., & Gans, S. J. (1983). Perception of the duration of rapid spectrum changes in speech and nonspeech signals. *Perception and Psychophysics*, *34*, 314–322.

Pisoni, D. B., & Lazarus, J. H. (1974). Categorical and noncategorical modes of speech perception along the voicing continuum. *Journal of the Acoustical Society of America*, *55*, 328–333.

Port, R. F., & Dalby, J. (1982). Consonant/vowel ratio as a cue for voicing in English. *Perception and Psychophysics*, *32*, 141–152.

Potter, R. K. (1950). Toward the specification of speech. *Journal of the Acoustical Society of America*, *22*, 807.

R Development Core Team. (2011). R: A Language and Environment for Statistical Computing [Computer software manual]. Vienna, Austria.

Repp, B. H. (1981). On levels of description in speech research. *Journal of the Acoustical Society of America*, *69*, 1462–1464.

Repp, B. H. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological Bulletin*, *92*, 81–110.

Repp, B. H., & Lin, H. (1991). Effects of preceding context on voice-onset-time category boundary. *Journal of Experimental Psychology: Human Perception and Performance*, *17*, 289–302.

Repp, B. H., & Mann, V. A. (1981). Perceptual assessment of fricative-stop coarticulation. *Journal of the Acoustic Society of America*, *69*, 1154–1163.

Ryalls, J., Zipprer, A., & Baldauff, P. (1997). A preliminary investigation of the effects of gender and race on voice onset time. *Journal of Speech, Language, and Hearing Research*, *40*, 642–645.

Salverda, A. P., Dahan, D., & McQueen, J. M. (2003). The role of prosodic boundaries in the resolution of lexical embedding in speech comprehension. *Cognition*, *90*, 51–89.

Sharma, A., & Dorman, M. F. (1999). Cortical auditory evoked potential correlates of categorical perception of voice-onset time. *Journal of the Acoustical Society of America*, *106*, 1078–1083.

Sharma, A., & Dorman, M. F. (2000). Neurophysiologic correlates of cross-language phonetic perception. *Journal of the Acoustical Society of America*, *107*, 2697–2703.

Sharma, A., Marsh, C. M., & Dorman, M. F. (2000). Relationship between N1 evoked potential morphology and the perception of voicing. *Journal of the Acoustical Society of America, 108*, 3030–5.

Shinn, P. C., Blumstein, S. E., & Jongman, A. (1985). Limitations of context conditioned effects in the perception of [b] and [w]. *Perception and Psychophysics, 38*, 397–407.

Smits, R. (2001a). Evidence for Hierarchical Categorization of Coarticulated Phonemes. *Journal of Experimental Psychology: Human Perception and Performance, 27*, 1145–1162.

Smits, R. (2001b). Hierarchical categorization of coarticulated phonemes: A theoretical analysis. *Perception and Psychophysics, 63*, 1109–1139.

Soli, S. D. (1981). Second formants in fricatives. *Journal of the Acoustical Society of America, 69*, S15.

Spivey, M. (2007). *The Continuity of Mind.* New York: Oxford University Press.

Steinschneider, M., Schroeder, C., Arezzo, J., & Vaughanjr, H. (1994). Speech-evoked activity in primary auditory cortex: effects of voice onset time. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section, 92*, 30–43.

Steinschneider, M., Volkov, I. O., Noh, M. D., Garell, P. C., & Howard, M. A. (1999). Temporal encoding of the voice onset time phonetic parameter by field potentials recorded directly from human auditory cortex. *Journal of Neurophysiology, 82,* 2346–2357.

Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *Journal of the Acoustical Society of America, 111*, 1872–1891.

Stevens, K. N., & Blumstein, S. E. (1978). Invariant cues for place of articulation in stop consonants. *Journal of the Acoustical Society of America, 64*, 1358–1368.

Stevens, K. N., Blumstein, S. E., Glicksman, L., Burton, M., & Kurowski, K. (1992). Acoustic and perceptual characteristics of voicing in fricatives and fricative clusters. *Journal of the Acoustical Society of America, 91*, 2979–3000.

Stevens, K. N., House, A. S., & Paul, A. P. (1966). Acoustical description of syllabic nuclei: An interpretation in terms of a dynamic model of articulation. *Journal of the Acoustical Society of America, 40*, 123–132.

Stevens, K. N., & Keyser, S. J. (2010). Quantal theory, enhancement and overlap. *Journal of Phonetics, 38*, 10–19.

Strand, E., & Johnson, K. (1996). Gradient and visual speaker normalization in the perception of fricatives. In D. Gibbon (Ed.), *Natural Langauge Speech Processing and Speech Technology* (pp. 14–26). Berlin: Mouton de Gruyter.

Strange, W., Jenkins, J. J., & Johnson, T. L. (1983). Dynamic specification of coarticulated vowels. *Journal of the Acoustical Society of America, 74*, 695–705.

Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance, 7*, 1074–95.

Sussman, H. M., Fruchter, D., Hilbert, J., & Sirosh, J. (1998). Linear correlates in the speech signal: The orderly output constraint. *Behavioral and Brain Sciences*, *21*, 241–259; discussion 260–299.

Sussman, H. M., & Shore, J. (1996). Locus equations as phonetic descriptors of consonantal place of articulation. *Perception and Psychophysics*, *58*, 936–946.

Swartz, B. L. (1992). Gender difference in voice onset time. *Perceptual and Motor Skills*, *75*, 983–992.

Syrdal, A. K. (1985). Aspects of a model of the auditory representation of American English vowels. *Speech Communication*, *4*, 121–135.

Syrdal, A. K., & Gopal, H. S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *Journal of the Acoustical Society of America*, *79*, 1086–1100.

Tanenhaus, M. K., Knowlton, S. M. J., Eberhard, K. M., & Sedivy, J. E. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*, 632–634.

Toscano, J. C., & McMurray, B. (2008, November). Online processing of acoustic cues in speech perception: Comparing statistical and neural network models. Poster presented at the 156th Meeting of the Acoustical Society of America, Miami, FL.

Toscano, J. C., & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science*, *34*, 434–464.

Toscano, J. C., & McMurray, B. (2011a). *Cue integration and context effects in natural and synthetic speech: Online processing, categorization, and simulation results.* Manuscript submitted for publication.

Toscano, J. C., & McMurray, B. (2011b). *Effects of sentence rate, voice-onset time, and vowel length on listeners' voicing judgments.* Manuscript in preparation.

Toscano, J. C., McMurray, B., Dennhardt, J., & Luck, S. J. (2010). Continuous perception and graded categorization: Electrophysiological evidence for a linear relationship between the acoustic signal and perceptual encoding of speech. *Psychological Science*, *21*, 1532–1540.

Tremblay, K. L., Friesen, L., Martin, B. A., & Wright, R. (2003). Test-retest reliability of cortical evoked potentials using naturally produced speech sounds. *Ear and Hearing*, *24*, 225–232.

Turk, A., Nakai, S., & Sugahara, M. (2006). Acoustic Segment Durations in Prosodic Research: A Practical Guide. In S. Sudhoff et al. (Eds.), *Methods in Empirical Prosody Research* (pp. 1–28). Berlin: Walter de Gruyter.

Utman, J. A. (1998). Effects of local speaking rate context on the perception of voice-onset time in initial stop consonants. *Journal of the Acoustical Society of America*, *103*, 1640–1653.

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories form infant-directed speech. *Proceedings of the National Academy of Sciences*, *104*, 13273–13278.

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 1181–1186.

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 and Performance*, *36*, 1005–1015.

Wayland, S. C., Miller, J. L., & Volaitis, L. E. (1994). The influence of sentential speaking rate on the internal structure of phonetic categories. *Journal of the Acoustical Society of America*, *95*, 2694–2701.

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

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

Whiteside, S. P. (2001). Sex-specific fundamental and formant frequency patterns in a cross-sectional study. *Journal of the Acoustical Society of America*, *110*, 464.

Whiteside, S. P., & Irving, C. J. (1998). Speakers' sex differences in voice onset time : A study of isolated word production. *Perceptual and Motor Skills*, *86*, 651–654.

Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis.* New York: Springer.

Woldorff, M. G. (1993). Distortion of ERP averages due to overlap from temporally adjacent ERPs: Analysis and correction. *Psychophysiology*, *30*, 98–119.

Wood, C. C. (1971). Auditory evoked potentials during speech perception. *Science*, *173*, 1248–1251.

Yeni-Komshian, G. H., & Soli, S. D. (1981). Recognition of vowels from information in fricatives: perceptual evidence of fricative-vowel coarticulation. *Journal of the Acoustical Society of America*, *70*, 966–75.