UNIVERSITY OF CALIFORNIA

Los Angeles

Prosodic prominence in vowel perception and spoken language processing

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Linguistics

by

Jeremy Andrew Steffman

2020

ABSTRACT OF THE DISSERTATION

Prosodic prominence in vowel perception and spoken language processing

by

Jeremy Andrew Steffman
Doctor of Philosophy in Linguistics
University of California, Los Angeles, 2020
Professor Sun-Ah Jun, Chair

This dissertation tests how prosodic prominence mediates the listener's path from the speech signal to segmental categories in perception. Though prosody and segment can be taken to represent different components of phonological structure, the central idea pursued here is that in processing they must necessarily interact on the basis of the way they jointly shape acoustic information in the speech signal. The experiments contained in this dissertation accordingly address how prominence affects listeners' categorization of speech sounds, how different prominence-lending contexts impact perception, and how prominence information and segmental information are integrated in online processing. Perception of vowel contrasts is adopted as a test case, in light of the way in which vowel realizations are strengthened phonetically when prominent.

This dissertation finds that prominence-lending context shifts listeners' perception of vowel contrasts, cued by the first formant (F1) and second formant (F2), such that a more prominent, or "strengthened", realization of a vowel is expected by listeners when a vowel is contextually prominent. Two ways of lending prominence are tested: (1) a manipulation of accentual/phrasal prominence in which a target bears the nuclear accent in a phrase, or is unaccented following narrow focus, and (2) the presence/absence of glottalization immediately preceding an accented vowel-initial target word, where glottalization can be seen as a manipulation of localized, phonetic prominence cues (with accentual status invariant). The timecourse of listeners' integration of prominence information is tested using a visual

world eyetracking paradigm. This timecourse assessment finds that both phrasal prominence and glottalization shape listeners' early processing of formant cues, showing a pattern that could be characterized as immediate compensation for contextual influences on the vowel. However, these two manipulations differ in that the effect of phrasal prominence is overall delayed in relation to listeners' use of formant cues, and reaches its maximum later in processing. This outcome is consistent with recent proposals which demarcate phrasal prosodic influences as entering relatively late into processing. In comparison, glottalization shows a near-synchronous timecourse with formant cues in processing, indicating that localized prominence information is integrated rapidly. Comparison of these two effects shows that phonetic prominence information influences processing differently than prominence linked to a more global/phonological structure, though even in the case of phrasal prominence, early effects on formant processing arise as a function of relative phonetic prominence in relation to context.

The dissertation also tests how vowel contrasts varying in height, which are subject to different patterns of prominence strengthening, are perceived based on prominence. It is observed that the perceptual expectations for a prominent vowel vary based on vowel features. Specifically, high front vowels, when prominent, are expected to show more hyperarticulated realizations with peripheral F1 and F2 (lowered F1, raised F2). In contrast, non-high vowels are expected to show more open realizations with less peripheral F1 and F2 (raised F1, lowered F2). This result shows that listeners integrate vowel-specific (or, feature-specific) expectations for how a vowel is realized in a prominent context.

These insights are discussed in light of other recent findings which test the influence of prosodic boundaries in segmental perception, and in relation to a more general model of both pre- and post-lexical contextual influences in spoken language processing. The schematic architecture of a model of prominence and segmental processing called Multistage Assessment of Prominence in Processing (MAPP) is proposed.

The dissertation of Jeremy Andrew Steffman is approved.

Taehong Cho

Megha Sundara

Patricia Keating

Sun-Ah Jun, Committee Chair

University of California, Los Angeles

2020

for my parents

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

This dissertation would not have been possible without the help of many people. I owe a debt of gratitude to all of them for their advice, guidance and support.

First, I am incredibly grateful to Sun-Ah Jun for many things: for teaching me about prosody, for being an invested mentor and collaborator, and for pushing me to extend my ideas and think creatively about research. More generally, for always being supportive, kind, and full of sage advice and inspiration. Sun-Ah made my time in the program low-stress and intellectually rewarding - I can only hope to replicate a fraction of her wisdom as an advisor in my own future endeavors.

I am further grateful to my committee members Pat Keating, Megha Sundara and Taehong Cho for their essential input on the dissertation, for being nothing but supportive and encouraging, and for their mentorship throughout my time at UCLA. Going back in time a little bit, I am also grateful to Susan Lin for being an outstanding mentor to me as an undergraduate, and for first piquing my interest in phonetics and experimental methods which started me down this path.

I could not have completed this work without the help of excellent undergraduate research assistants throughout my graduate career, both in projects which led up to and inspired this dissertation work, and in data collection for the dissertation itself. For their help in various tasks and their enthusiasm I am grateful to Danielle Bagnas, Juliana Casparian, Bryan Gonzalez, Qingxia Guo, Yang Wang and Jae Weller.

I would also like to thank Henry Tehrani for technical help, and Adam Royer for generously lending his time to record (very repetitive) speech materials and offer his help and advice. For helpful discussion on these projects, using an eye tracker, and related work during the process of writing this dissertation I am grateful to Adam Chong, Marc Garellek, Hironori Katsuda, Sahyang Kim and Chie Nakamura, and to audiences at LabPhon 17, CUNY 33, WCCFL 38 and numerous UCLA phonetics seminars.

More generally, I am grateful to have found such a supportive environment in the UCLA

Phonetics Lab, which nourished my ideas and research agenda, and offered the resources and support that made this dissertation possible. I am also thankful to have learned so much from excellent teachers in the UCLA Department of Linguistics during my time in the program.

To all of my colleagues in the department, I am thankful for the chats, the thought-provoking conversations, the times spent at a coffee shop, the pomodoro study sessions, and for generally making UCLA an enjoyable place to be a student. To my other friends and family: thank you for being around and being supportive.

Finally, I cannot thank Marissa enough, for everything.

VITA

2018 M.A. (Linguistics)
University of California, Los Angeles

2016 B.A. (Linguistics and French)
University of California, Berkeley

PUBLICATIONS

Steffman, J. & Katsuda, H. (2020). Intonational structure influences perception of contrastive vowel length: The case of phrase-final lengthening in Tokyo Japanese. *Language and Speech*, 0023830920971842.

Steffman, J. & Jun, S.-A. (2019). Perceptual integration of pitch and duration: Prosodic and psychoacoustic influences in speech perception. *The Journal of the Acoustical Society of America*, 146(3), EL251– EL257.

Steffman, J. (2019). Phrase-final lengthening modulates listeners' perception of vowel duration as a cue to coda stop voicing. *The Journal of the Acoustical Society of America*, 145(6), EL560–EL566.

Steffman, J. (2019). Intonational structure mediates speech rate normalization in the perception of segmental categories. *Journal of Phonetics*, 74, 114–129.

Steffman, J. (2018). Nominal inflection in the Safané dialect of Dafing: Ternary quantity contrasts and morphologically conditioned phonology. *Journal of West African Languages*, 45(1), 89–124.

CHAPTER 1

Introduction

1.1 What this dissertation is about

This dissertation is about how listeners understand spoken language. Two key parts of this task are (1) determining contrastive segmental categories which convey lexical distinctions, and (2) determining prosodic categories which convey various other pieces of information, such as phrasal grouping, prominence relations, information structure, and so on (e.g., Christophe, Peperkamp, Pallier, Block, & Mehler, 2004; Cutler, Dahan, & Van Donselaar, 1997; Mitterer, Kim, & Cho, 2019; Salverda et al., 2007; Schafer, 1997). The experiments in this dissertation test how these two components of spoken language processing interact, and more specifically, how prosodic prominence mediates listeners' perception of vowel contrasts.[1] Though prosody and segment can be construed as two separate types of linguistic structure in spoken language, this dissertation shows that the listener's path from the speech stream to a segmental and prosodic parse involves considerable interaction between these two domains on the basis of how they, together, shape acoustic information in the speech signal.

Important questions about the effects of prosody on the perception of speech segments, and the processing responsible for these effects, remain to be answered. A large part of what we do know comes from experiments testing the influence of prosodic boundaries. The role of prosodic prominence is less studied, and as such, tests of prominence effects in seg-

[1]Traditionally, models of spoken word recognition focus on segmental information (McClelland & Elman, 1986; Norris, 1999), including the listener's task of parsing the speech stream into contrastive segmental categories. Effects of prosody (including word-level prosody) are usually studied in the context of word segmentation and lexical access (Brown, Salverda, Dilley, & Tanenhaus, 2015; Dilley, Mattys, & Vinke, 2010; Salverda et al., 2007), or as they relate to syntactic and other strictly post-lexical effects in speech comprehension (Cutler et al., 1997; Price, Ostendorf, Shattuck-Hufnagel, & Fong, 1991; Wagner & Crivellaro, 2010).

mental perception offer an avenue for answering various empirical and theoretical questions. Throughout the experiments in this dissertation we will explore how listeners' perception of formants (cuing vowel contrasts) changes based on how prosodically prominent a given sound is. We will also explore if different prominence-lending contexts generate similar perceptual effects and will test how contextual prominence-lending information is integrated with formant cues in online processing.[2] Results from eyetracking experiments will be used to inform a theory of how prominence relates to segmental processing, and how this compares to existing models. In so doing, we will touch on various theoretical questions related to domain-generality in speech perception, abstraction and retention of detail in processing, and stages and information flow in lexical access.

1.2 The phonetic encoding of prosodic structure

Why should phrasal prosody (i.e. grouping above the word, phrasal prominence, intonational tunes, and so on) matter for perception of segmental contrasts? The relevance of phrasal prosody in this domain becomes apparent if we consider how segmental realizations vary based on prosodic factors, and how listeners are influenced by contextual variation in speech. These points are discussed below.

1.2.1 Prosody

This dissertation conceives of prosody as an abstract "raw organizational structure" (Beckman, 1996, p 19). A useful definition is given by Cho (2016, p 122):

> Under this structural view [...] the term prosody no longer refers merely to lower-order suprasegmental features such as pitch, duration, and amplitude, but it embraces an abstract notion of a higher-order grammatical structure definable

[2]Here and throughout this dissertation, "prominence-lending" is used to refer to properties which signal a unit of speech as prominent, i.e. as synonymous with a term like prominence-cuing. This differs from the more specific definition of prominence-lending used in the IPO tradition (e.g., 't Hart, Collier, & Cohen, 1990). See e.g., Ladd (2008, p 54) for discussion.

> as "a hierarchically organized structure of phonologically defined constituents and heads" (Beckman, 1996, p 19). It therefore provides a frame for articulation with two functions: a delimitative function regarding how smaller phonological units or prosodic constituents (phonemes, syllables) are grouped together to form a larger prosodic constituent (a prosodic word or a phrase) and a culminative function regarding which of the prosodic constituents in the utterance should be the head of the phrase.[3]

In the more traditional definition alluded to in the quote above, prosody is seen as linguistically meaningful patterns of duration, intensity, and f0, corresponding to the perception of duration/length, loudness and pitch.[4] Prosody is "overlaid" on stretches of speech that span multiple segments, hence the term *suprasegmentals*, which has been used as a synonym in some cases (see e.g., Fletcher, 2010).

To contrast the traditional conception of prosody as suprasegmentals with the organizational/structural view adopted here, we can consider some problems with the idea that prosody is confined to variation in duration, pitch and loudness. In pointing out one such problem, Lehiste (1970) remarks that segments differ intrinsically in these measures. For example, vowels vary in all three of these properties as function of features like vowel height: high vowels generally have higher pitch and shorter duration as compared to low vowels (e.g., Hillenbrand, Getty, Clark, & Wheeler, 1995; Lehiste & Peterson, 1959; Peterson & Barney, 1952; Peterson & Lehiste, 1960). Pitch and duration also vary systematically as a function of voicing contrasts (e.g., Chen, 1970; Löfqvist, Baer, McGarr, & Story, 1989; Whalen, Abramson, Lisker, & Mody, 1993). These dependencies are clearly linguistically meaningful in the sense that listeners exploit patterns of variation in pitch and duration in their perception of segmental material (e.g., Chuang & Wang, 1976; Raphael, 1972; Shultz, Francis, & Llanos, 2012; Steffman & Jun, 2019; Yu, Lee, & Lee, 2014). As such, the idea

[3]For further descriptions of this view of prosodic structure see e.g., Beckman (1996); Beckman, Edwards, and Fletcher (1992); Cho (2015, 2016); de Jong, Beckman, and Edwards (1993); Fletcher (2010).

[4]The term pitch is used predominantly in this dissertation (in lieu of f0) as the perceptual definition is most relevant.

that pitch (for example) is only a prosodic property, or that pitch *is* prosody is irreconcilable with the fact that segments show intrinsic and meaningful variation in pitch. We can imagine further issues with this definition in languages with lexical tone contrasts, or even in languages without lexical tone where pitch serves as cue to e.g., lexically contrastive metrical stress (Zahner, Kutscheid, & Braun, 2019). A definition of pitch as prosody that conflates lexically contrastive uses of pitch with post-lexical or phrasal pitch patterns (e.g., intonational tunes) misses an important distinction.

On the other hand, acoustic features in the speech signal besides f0, duration and intensity vary systematically as a function of prosodic organization (e.g., Keating, 2006; Lehiste, 1970). Put differently, these properties vary in a patterned way that is compatible with the notion of a hierarchically organized prosodic structure (i.e. well beyond immediate segmental context), and *co-vary* systematically with other acoustic features such as f0, duration and intensity. This includes properties which might in a traditional view be seen as purely "segmental", such as voice onset time (VOT) and formant structure, a point that is discussed in detail below.

If the boundary between segment and prosody cannot be drawn in a satisfying way on the basis of acoustic medium, an alternative is to posit prosody as a more abstract, organizational entity in its own right, which structures the speech signal in a systematic way. It is worth remarking here that a structural definition of prosody does not mean that f0, duration and intensity are not important facets of prosodic organization, as noted by Fletcher (2010, p 528):

> The two uses of the term prosody [...] as either suprasegmentals or abstract hierarchical phonological structure, are not completely unrelated because phonetic parameters like f0, intensity and duration (the "classic" suprasegmental parameters according to most phonetics textbooks) [...] contribute to the signaling of different aspects of prosodic structure.

As noted in the previous quote from Cho (2016), it is common in theories of prosodic organization to posit two primary functions of prosody. The delimitative function groups

linguistic units, and correlates with (though is independent from) syntactic structure. The culminitive function could also be described as a prominence-marking function, where the prosodic structure assigns particular structural positions to be prominent within a given domain. This dissertation adopts this view in making a general distinction between boundary-marking and prominence-marking functions of prosodic organization.

There are some noteworthy consequences of this theory of prosody as organizational structure outlined above. First, in de-coupling prosody from a particular set of acoustic properties, we can explore how other, non-"suprasegmental" acoustic features are used to mark prosodic organization. The data outlined below in Section 1.2.2 indicates that this a necessary step, in light of just how much various acoustic and articulatory parameters are influenced in a systematic way by prosody.

Framing prosody as an abstract organizational entity has also placed it front and center in models of speech production planning (Keating & Shattuck-Hufnagel, 2002; Krivokapić, 2012; Shattuck-Hufnagel, 2000) given that a prosodic frame for a planned utterance will influence the pitch pattern over a phrase and phonological alternations, down to the timing and amplitude of how individual segments are articulated. This must include considerable look-ahead in planning prosodic structure, given the scope of e.g., prominence placement and phrasing effects that extend forward from the start of a planned utterance (see Keating & Shattuck-Hufnagel, 2002 for an overview). A central role for prosody in speech production might suggest a correspondingly central role in speech comprehension (Cho, McQueen, & Cox, 2007; Cutler et al., 1997), in line with the idea that prosody is "[...] a complex grammatical structure that must be parsed in its own right." (Beckman, 1996, p 64). This is discussed further below.

As is apparent from the theory outlined above, we might expect to see effects of prosodic organization playing out in detailed ways in the phonetic properties of speech segments. In what follows, some ways in which prosody is encoded in speech articulations and acoustics are outlined, after which implications for speech perception are discussed.

1.2.2 Phonetic encoding

A large body of research shows that prosody influences the phonetic realization of segments, conceptualized as the phonetic encoding of prosodic structure (Keating, 2006). The general idea is that phrasal prosodic organization serves as a frame for the articulation of segments, fine-tuning the timing and amplitude of segmental articulations as stated above. Following the distinction made between the prominence-marking and boundary-marking function of prosodic organization, some ways in which both are encoded phonetically are outlined below.

First consider two well-known examples related to prosodic boundaries: *domain-initial strengthening* and *pre-boundary lengthening*. Both of these effects systematically modulate how segments are realized, especially in terms of their temporal structure (though spectral structure can also be impacted by boundaries, see e.g., Georgeton, Antolík, & Fougeron, 2016).

In domain-initial strengthening, segmental articulations are realized in a "stronger" fashion, with increased articulatory contact, longer closure duration etc., when at the beginning of a prosodic domain as compared to in the middle of a prosodic domain (e.g., Cho, 2015, 2016; Fougeron & Keating, 1997; Keating, Cho, Fougeron, & Hsu, 2004). Initial strengthening can manifest as temporal expansion of gestures, though this is notably dependent on the properties of a given segment that is undergoing strengthening (Cho & Keating, 2001; Cho, Kim, & Kim, 2017; Fougeron, 2001). One key finding in this literature is that degree of strengthening generally maps hierarchically onto prosodic domains whereby higher-level domains show increased strengthening. For example, Fougeron and Keating (1997), found that linguopalatal contact was greater initial to an intonational phrase (IP) than an intermediate phrase (ip) than a word.[5] One well-studied acoustic consequence of initial strengthening is manifested in voice onset time (VOT). In aspirated stops, VOT is longer at the beginning of an IP, as compared to being IP-medial and varies as a function of hierarchical prosodic organization (Cho & Keating, 2001, 2009; Jun, 1996). As an illustration, VOT data from Cho

[5]Here terminology from the prosodic hierarchy described in Beckman and Pierrehumbert (1986) and Pierrehumbert (1980) is adopted, in keeping with much of the literature on domain-initial strengthening. This terminology is further used throughout the dissertation.

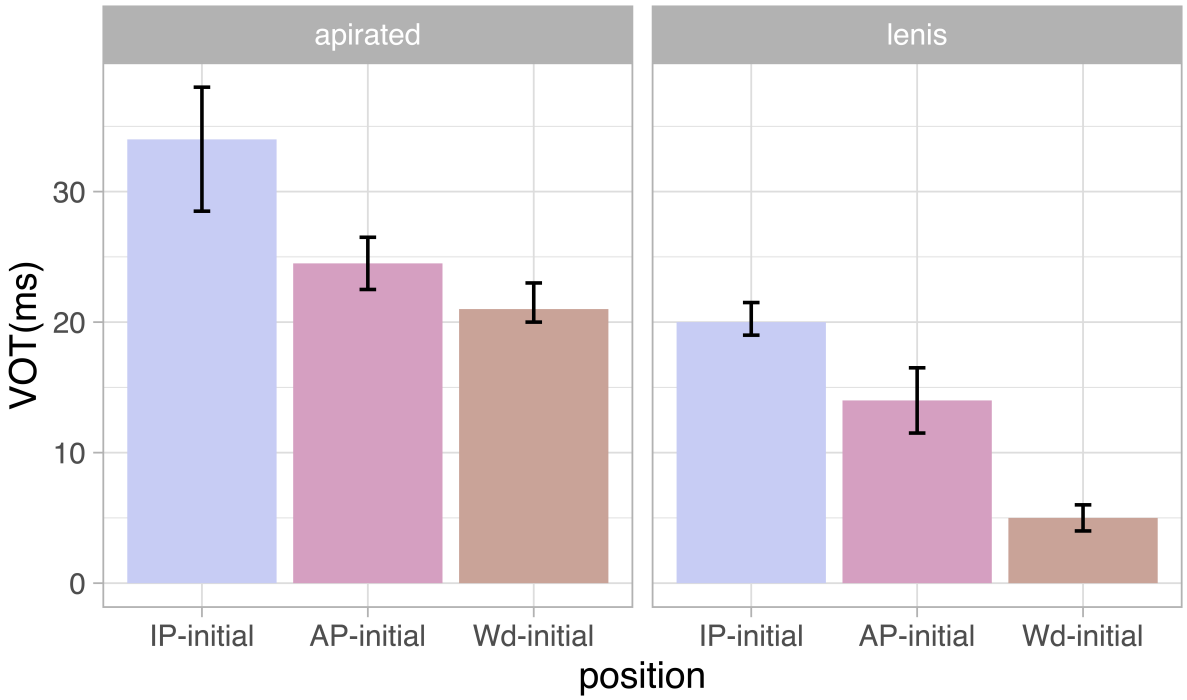

Figure 1.1: VOT for Korean aspirated /t^h/ (left panel), and lenis /t/ (right panel), in different prosodic positions, adapted from Cho and Keating (2001). Error bars show 97% confidence intervals.

and Keating (2001) is shown in Figure 1.1 for the Korean stops /t^h/ and /t/. Following Jun (1996, 1998), three domains are shown: the intonational phrase (IP), the accentual phrase (AP, a smaller phrase with a domain slightly larger than a word), and the word.

Figure 1.1 illustrates how variation in VOT for these speakers systematically patterns as a function of segmental/featural specification, and simultaneously as a function of prosodic organization.[6] Note that, overall, aspirated /t^h/ has longer VOT than lenis /t/. Additionally, at higher prosodic domains VOT is longer, and systematically decreases moving down the prosodic hierarchy. This prosodically-induced variation generates some overlap in segmental categories: word-initial /t^h/ has VOT that is comparable to IP-initial /t/, while

[6]Note that the Korean speakers who generated the data are older speakers who do not display the merger in VOT for these categories which is evident in younger speakers (Choi, Kim, & Cho, 2020; Kang, 2014). It also worth noting that this contrast is additionally conveyed by pitch on the following vowel (see e.g., Jun, 1996, 1998), such that this particular contrast isn't neutralized as a function of prosodically-driven variation.

IP-initial VOT in /t^{h}/ is much longer. An idea proposed in the literature (discussed below) is accordingly that listeners would benefit from determining how prosody has shaped a cue's value (Cho et al., 2007; Kim & Cho, 2013). For example, IP-initial VOT for /t/, if not reconciled with the fact that is is IP-initial, might be misleadingly long to listeners, falling more in the range of /t^{h}/. Put differently, after hearing a stop with 20 ms of VOT (given the distributions in Figure 1.1), the listener will not know whether this should map to /t/ or /t^{h}/ without also knowing whether it is IP-initial or word-initial. The reconciliation of prosodic context with a cue value would be functionally useful to listeners in this sense.

More generally, VOT is a strong cue to laryngeal contrasts cross-linguistically, where, for example, longer VOT cues a voiceless stop (in e.g., English) or an aspirated stop (in e.g., Thai, or Korean). In this sense, and in its traditional framing, VOT is a "segmental cue", that is, it is clearly important for conveying contrasts at the segmental level (Abramson, 1976; Abramson & Whalen, 2017; Lisker & Abramson, 1964, 1970). Nevertheless, as Figure 1.1 clearly shows, a given duration of VOT will be shaped both by laryngeal specifications in a segment, and by prosodic phrasing (among other things). As stated above, listeners would benefit from disentangling the segmental and prosodic contribution to VOT duration, both in determining what laryngeal category it cues, and for more upstream processes like word segmentation (Cho et al., 2007). This is discussed below in Section 1.4.2.

Complementary to initial strengthening, pre-boundary lengthening (also called phrase-final lengthening), is another well-established way in which phrasal boundaries impact segmental realizations (Klatt, 1975, 1976; Turk & Shattuck-Hufnagel, 2007; Wightman, Shattuck-Hufnagel, Ostendorf, & Price, 1992). As the name suggests, this refers to the temporal expansion (or articulatory slowing of speech gestures) for linguistic units preceding the boundary of some prosodic domain (e.g., Byrd & Saltzman, 2003; Cho, 2015, 2016; Krivokapić & Byrd, 2012). In similar fashion to the example of VOT and initial strengthening, we can consider that the value of a given durational cue (e.g., vowel duration in a language with contrastive vowel length) will vary both as a function of its segmental/featural specification (short or long) and prosodic position (pre-boundary, or not) (Nakai, Kunnari, Turk, Suomi, & Ylitalo, 2009; Shepherd, 2008). In order to determine the segmental specification of a given vowel

in this case, listeners would accordingly benefit from factoring in how it has been impacted by prosody.[7] Durational cues more generally, e.g., vowel duration as a cue to coda voicing (Raphael, 1972) would be similarly relevant here. Pre-boundary lengthening is thus another systematic way in which segments are modulated by prosody, with possible perceptual relevance to listeners (discussed below).

Given the patterns overviewed thus far, an unanswered question is *why* effects such as these occur. One common argument in the literature is that the prosodic fine-tuning of segmental realization is not simply a consequence of bio-mechanical constraints (e.g., longer VOT as the by-product of generally more forceful articulations domain-initially), but instead that these patterns serve a linguistic function. As such, they are controlled by the speaker as part of the phonetic grammar of a language (see e.g., Cho, 2015, 2016; Cho & Keating, 2001; Garellek, 2013; Keating, 2006; Keating et al., 2004). What linguistically useful functions might these boundary-driven effects serve? Following the theory of prosodic structure sketched above, prosodic boundaries are supposed to serve a delimitative function. One previously argued function of domain-initial strengthening is accordingly *syntagmatic contrast enhancement*. Here the term syntagmatic refers to the relationship between neighboring linguistic units in the speech stream (e.g., segments). Syntagmatic enhancement helps set units apart in a sequence, making them more distinct from surrounding material. Consider the case of domain-initial lengthening of VOT in an initial C^hV. Longer VOT (and oral closure duration; Cho & Keating, 2009) sets the initial consonant apart from preceding material (cuing a boundary) and can simultaneously be seen as helping perceptually separate the consonant from the following vowel in enhancing its consonantal features (Cho, 2015).

One reason to think of these effects as enhancement of contrasts between adjacent units, as opposed to simply making a more clearly articulated segment, is the existence of cases where domain-initial strengthening seems to come at the "expense" of segment-internal features. Consider a few examples. For nasals, the duration, nasal energy, and scope of coarticulation with a following vowel are all lessened domain-initially (Cho & Keating, 2009; Cho et

[7]Nakai et al. (2009) found overlap in short and long Finnish vowels as a function of pre-boundary lengthening, suggesting clearly that computing phrasal position would be helpful in this case.

al., 2017; Fougeron, 2001), reducing the sonority and general salience of the segment. As another example, VOT for English voiced stops, where these "voiced" stops are usually voiceless with short-lag VOT in word-initial position (e.g., Davidson, 2016), is also lengthened domain-initially (Kim, Kim, & Cho, 2018). Additionally, the English fricatives /ð/ and /θ/ are stopped more frequently in domain-initial position (Melguy, 2018). What these cases have in common is that segments are strengthened in a way that diminishes their intrinsic featural specification (e.g., [+sonorant] for nasals, [+voice] for voiced stops [+continuant] for fricatives), but these changes consistently help set the segment apart from a following vowel and signal a break with preceding material.

Another strengthening effect argued to occur at prosodic boundaries is *paradigmatic contrast enhancement*. The term paradigmatic, as used here, refers to how linguistic units relate to one another independent of context, for example, the relationship between vowel phonemes in a language or more generally the featural relationships between classes of sounds (e.g., voicing). Paradigmatic enhancement accordingly refers to changes that make a given unit distinct from others in a language system. In the literature, paradigmatic enhancement generally refers to the enhancement of acoustic properties that maximize phonemic contrasts in prosodically strong positions. Though the literature suggests that, at least in American English, paradigmatic enhancement effects are more related to prominence (Cho, 2015; de Jong, 1991, 1995), prosodic boundaries have also been shown to play a role, in line with the idea that domain-initial position is privileged in terms of contrast maintenance (Beckman, 1998; Steriade, 1997). For example, Georgeton et al. (2016) found that French vowels in IP-initial position showed acoustic and articulatory changes that generally maximized their distinctiveness from one another, enhancing the contrasts between vowels in the language (as opposed to between a vowel and neighboring material). Guitard-Ivent, Chignoli, Fougeron, and Georgeton (2019) also found that when acoustic data was given to a trained classifier, vowel classification accuracy was overall improved in IP-initial (as compared to word-initial) position. Cho (2004) also shows that in English, IP-initial vowels are resistant to coarticulatory influences, lining up with this idea. Another example that suggests paradigmatic contrast enhancement at prosodic boundaries comes from Cho and Jun (2000). Recall that

Korean aspirated and lenis stops showed longer VOT when initial to higher-level prosodic domains, as shown in Figure 1.1. Cho and Jun, however, found that Korean fortis stops showed slightly *shorter* VOT at higher prosodic domains, the opposite pattern observed for the other laryngeal categories. This suggests that speakers are enhancing the laryngeal features of the three-way laryngeal contrast domain-initially by implementing short VOT for fortis stops (see also Cho & Keating, 2001, and Cho & McQueen, 2005 for a comparable case in Dutch).

Much of the research on prosodic boundaries that has been framed around contrast enhancement has dealt with domain-initial strengthening. Domain-final effects are generally not seen as strengthening in the same way, as domain final position is typically construed as being weaker prosodically (Cho, 2016). Nevertheless, domain-final effects have also been shown to be equally sensitive to linguistic factors. If pre-boundary lengthening was only a general physiological slowing of articulators before the cessation of speech (Berkovits, 1993; Krakow, Bell-Berti, & Wang, 1995; Lindblom, 1968), we would expect it to be implemented in the same way across languages. However, the domain and scope of lengthening varies cross-linguistically, in e.g., interacting with the prominence system of a language (Katsika, 2016; Seo, Kim, Kubozono, & Cho, 2019; Turk & Shattuck-Hufnagel, 2007) and other phonological factors like contrastive length distinctions, and syllabic and moraic structure (Nakai et al., 2009; Seo et al., 2019; Shepherd, 2008).

The data in this section thus suggests that boundary-driven fine-tuning of segmental realizations constitutes a part of the speaker's phonetic grammar, being sensitive to various other components of linguistic structure. We can expect that listeners would further benefit from awareness of these patterns in segmental processing as stated above.[8] With the data reviewed so far we have seen how prosodic (boundary) and segmental specifications can combine in structuring how segments are realized. Next, we can explore how prosodic *prominence* might similarly shape segmental detail in speech.

[8]It can be noted that there is clear support for the influence of prosodic phrasing in lexical processing, where both domain-initial and domain-final patterns facilitate word segmentation and lexical access (Cho et al., 2007; Endress & Hauser, 2010; Kim, 2004; Kim & Cho, 2009; White, Benavides-Varela, & Mády, 2020).

1.3 Prominence strengthening

The relationship between prosodic structure and segmental detail outlined thus far has dealt with prosodic boundaries. Perhaps unsurprisingly, prosodic prominence has also been shown to shape how segments are realized. One well-documented effect is *prominence strengthening*. Strengthening is used here as a cover term to refer to the observed temporal and spatial modulations associated with articulations when they are produced as prominent by speakers. Prominence strengthening could in this way be defined as the way in which speakers change the realization of a segment to mark it as prominent. In this section, a working definition of prominence is given, and the ways in which prominence strengthening influences segmental realizations are outlined.

1.3.1 Prosodic prominence

In describing what sorts of segmental modulations constitute prominence strengthening, we first need to define what constitutes prominence. This is not an easy task, and the notion of prominence as a linguistic phenomenon is highly complex, with no simple characterization (Baumann & Cangemi, 2020; Baumann & Winter, 2018; Bishop, Kuo, & Kim, 2020). As Baumann and Cangemi (2020, p 1) state:

> Few concepts in phonetics and phonology research are as widely used and as vaguely defined as is the notion of prominence. At the crossroads of signal and structure, of stress and accent, and of production and perception, the notion of prominence has received a wide number of contradicting or unspecific definitions.

In what follows, a particular view of prosodic prominence is outlined, with reference to phonological organization and phonetics. We will then turn to the supposed function of prosodic prominence and describe prominence strengthening in this light. This will entail consideration of both phonological/categorical prominence distinctions, as well as phonetic/gradient prominence-marking cues. The idea that prosodic features should be considered both in terms of abstract/symbolic entities and phonetic parameters is of course not

new. Consider for example the quote from Cole and Shattuck-Hufnagel (2016, p 5):

> [...] prosody can be decomposed into two components, with discrete prosodic elements that encode structural and meaning relations among linguistic units such as words and phrases, phonetic cues that are bundled in systematic and potentially language-specific patterns in the acoustic speech signal, and different cue patterns in different contexts.[9]

First, in regards to the discrete and structuralcomponent of the quote above, one obvious determinant of prominence in a language is the phonological organization of prosodic structure. At the phonological level, prominence in a language like American English can be described as docking on metrically strong syllables (Beckman & Pierrehumbert, 1986; Hayes, 1995; Liberman & Prince, 1977; Nespor & Vogel, 2007; Pierrehumbert, 1980), and further determined by information and discourse structure, e.g., contrast, given-ness, etc. (Bolinger, 1958, 1961; Hirschberg & Pierrehumbert, 1986; Schwarzschild, 1999; Selkirk, 1995; Truckenbrodt, 1995). In this sense, prominence is configurational, related to both metrical and phrasal structure. For example, in a typical declarative utterance in American English, the most prominent syllable in a phrase will be the last accented syllable (e.g., Chomsky & Halle, 1968; Pierrehumbert, 1980), where tonal accents link to metrically strong positions. This prominence is thus "structural" in that it depends on metrical structure and patterns, prosodic phrasing, and the position of a prominent syllable within a phrase.

At the same time, prominence (and listeners' perception of it) integrates diverse and varied pieces of information, including features not directly related to the speech signal itself such as word frequency, information structural context, and part of speech (Baumann & Winter, 2018; Bishop, 2012, 2017; Calhoun, 2007; Cole et al., 2019).[10] This makes characterizing the concept of prominence in a complete manner a difficult task, as suggested by Baumann and Cangemi (2020) above (see also Wagner et al., 2015). Even if we restrict ourselves to

[9]See also Cangemi and Grice (2016); Grice, Ritter, Niemann, and Roettger (2017) for related ideas.

[10]Prominence is also conveyed visually by e.g., beat gestures and facial expressions (Bosker & Peeters, 2020; Krivokapić, 2014; Swerts & Krahmer, 2008).

information in the speech signal, defining prominence is complicated. The most general definitions have accordingly been said to be the most successful (Baumann & Cangemi, 2020; Wagner et al., 2015). One such general definition, which is adopted implicitly or explicitly in various studies, is given by Terken and Hermes (2000, p 89):

> We say that a linguistic entity is prosodically prominent when it stands out from its environment by virtue of its prosodic characteristics. That is, we define prominence as a property of a linguistic entity relative to an entity or a set of entities in its environment. Although the definition is cast in relative terms, it includes monosyllabic utterances, because they stand out from silence.

Following this definition, various phonetic properties have been shown to lend prominence in a granular fashion, i.e. beyond categorical distinctions encoded in models such as Pierrehumbert (1980). One often-employed test for how a certain cue, or piece of information (e.g., part of speech), signals prominence is to test experimentally if it is judged to lend prominence in a task where participants listen to speech. This can be done by providing listeners with a scale (e.g., a Likert-style scale from 1-5) and asking them to give numerical prominence ratings to words (Bishop, 2012). Another widely used method is so-called Rapid Prosody Transcription (RPT; Cole, Mo, & Hasegawa-Johnson, 2010; Cole, Shattuck-Hufnagel, & Mo, 2010; Mo, 2011). In an RPT task, participants are instructed to designate where they hear prominence, and where they hear boundaries, as they listen to a speech. For prominence, a continuous P(rominence)-score for each word in an annotated speech sample is calculated as the proportion of times listeners indicated it as prominent. With many participants completing this task, a P-score thus gives a granular estimate of how often a word is perceived as prominent. Next, the extent to which certain parameters predict P-scores can be assessed as a method of testing their role in marking prominence. In this sense, the question of "what properties lend prominence?", outside of phonological models of prosody and intonation, is addressed primarily with experimental/perceptual data.

As it pertains to acoustic properties in the signal, we can take this sort of data to answer

the question of what, phonetically, lends prominence in speech.[11] Three factors that have, unsurprisingly, been shown to lend prominence in this regard are increases in duration, pitch, and loudness (Baumann & Winter, 2018; Cole et al., 2019; Mo, 2008, 2011). P-scores increase in a continuous fashion as pitch and duration increase, showing that listeners rely on fine-grained detail in prominence perception, not simply on an abstract, or structural, indication of prominence (see also e.g., Fant & Kruckenberg, 1989; Katz & Selkirk, 2011; Mücke & Grice, 2014 for related ideas).[12] Clear evidence for this comes from cases where, even within a pitch accent category, phonetic parameters influence listeners' perception of prominence and its linguistic functions (e.g., focus marking), as shown by Bishop et al. (2020) and Grice et al. (2017).

These two different views of prosodic prominence could therefore be defined in structural terms (i.e. is a word accented phonologically?), and in phonetic terms (i.e. does this property lend prominence?). These will be referred to as phonological prominence, and phonetic prominence, respectively. There is an obvious relationship between the two: a structurally/phonologically prominent word is likely to be phonetically prominent and vice versa, but the difference is crucially in the level of structure and detail that is being considered. Because prominence perception is multi-faceted and granular, it is clear at the very least that we should look beyond purely structural descriptions of prominence to understand how it functions, particularly in speech perception and processing. With this distinction, and this working definition of prominence, we can now turn to how segmental features encode prominence, analogous to the phonetic encoding of boundaries described above.

[11]It is worth noting here too that structural aspects of prominence are also captured by P-scores. For example, Cole et al. (2019) found that nuclear accented words are perceived reliably as more prominent than pre-nuclear accented words, which are perceived as more prominent than unaccented words, as predicted by models such as that in Pierrehumbert (1980).

[12]In comparison, prosodic boundaries are generally assumed to form discrete categories, encoding different levels of phrasing in a prosodic hierarchy (e.g., Beckman & Pierrehumbert, 1986; Selkirk, 1995). There is empirical support for this notion (Carlson, Clifton, & Frazier, 2001), but it is also not uncontroversial. Perceived boundary varies as a function of continuous variables (Mo, 2011), and boundary perception as indexed by a syntactic parse or scalar rating can be gradient as well (Krivokapić & Byrd, 2012; Wagner & Crivellaro, 2010), though Krivokapić and Byrd (2012) argue that gradient boundary perception could arise from recursive prosodic phrasing, i.e. as a reflex of layered structural organization.

1.3.2 Prominence strengthening as contrast enhancement

As with prosodic boundaries, the effects of prominence on segmental realizations have often been framed in terms of paradigmatic and syntagmatic enhancement. However, as mentioned above, prominence is often seen as having a tighter link to paradigmatic enhancement effects (Cho, 2015; de Jong, 1995). Several examples of prominence strengthening effects in consonants are surveyed below.

For nasals, phrasal prominence leads to increased duration (Cho et al., 2017; Fletcher, Stoakes, Loakes, & Singer, 2015) and nasal energy (Cho & Keating, 2009), both of which can be viewed as enhancing the nasal's inherent features (this differs markedly from domain-initial strengthening for nasals where they are shortened, as discussed in Section 1.2.2 above).

Prominence also has been found to enhance voicing contrasts for stops. For example, Kim, Kim, and Cho (2018) found that word-initial English stop voicing contrasts are enhanced when accented. Both phonologically voiced and voiceless stops showed increases in VOT when accented (where "voiced" stops are voiceless with short-lag VOT). However, the degree of VOT lengthening was observed to be substantially greater for voiceless aspirated stops, maximizing the phonetic difference between the stop categories under accent (unlike the minimization of the voicing contrast observed with domain-initial strengthening, discussed in Section 1.2.2). Cole, Kim, Choi, and Hasegawa-Johnson (2007) further found that some speakers shorten VOT in word initial "voiced" stops, and that prominence additionally led to larger differences in closure duration, another cue to the voicing contrast. Another clear case of prominence enhancing stop features was found in Dutch, where Cho and McQueen (2005) observed that speakers produced unaspirated /t/ with shorter VOT when prominent, enhancing its unaspirated featural status ([-spread glottis]).

English fricatives are also lengthened and produced with increased noise when prominent (Cho, Lee, & Kim, 2014; Silbert & de Jong, 2008), though interestingly Silbert and de Jong (2008) do not find that speakers selectively enhance spectral properties that would strengthen contrasts between different fricative categories. However, Chuang and Fon (2010) found that speakers of Taiwanese Mandarin enhanced the contrast between dental and retroflex fricatives

and affricates under prominence, showing more strongly differentiated centroid frequencies.

What these effects have in common is their strengthening of acoustic properties that would help listeners discriminate contrasts or more generally better perceive the unit that is undergoing strengthening. As such, they serve the function of paradigmatic contrast enhancement. Findings such as these support the idea that prominent material in the speech signal is informationally rich, and as such it should maximally intelligible and discriminable, following, e.g., Baumann and Cangemi (2020); Ladd (2008).

1.3.3 Prominence strengthening in vowels

Most relevant to the experiments in this dissertation, vowels are also subject to prominence strengthening, with effects that are taken to reflect both paradigmatic and syntagmatic enhancement. For example, Cho et al. (2017) found that prominent vowels in English showed substantially reduced coarticulation with both preceding and following nasal consonants, measured acoustically. This could be seen as enhancing both syntagmatic and paradigmatic contrast. On one hand, reduced coarticulation more cleanly separates nasal and vowel, facilitating segmentation. On the other, reduced nasality in the vowel enhances its inherent acoustic properties and maximizes its contrast with other vowel phonemes.

One well-documented pattern of prominence strengthening for vowels in the literature is *sonority expansion* (Beckman et al., 1992; de Jong et al., 1993; Erickson, 2002). Here sonority is defined in phonetic/articulatory terms, as "the overall openness of the vocal tract or the impedance looking forward from the glottis" (Silverman & Pierrehumbert, 1990, p 75). A more sonorous articulation is accordingly one which is produced with increased amplitude of jaw movement and other articulatory adjustments that allow more energy to radiate from the mouth. Sonority-expanding gestures make a vowel articulation more acoustically prominent (louder, longer etc.), and have been described as enhancing its "sonority features" (de Jong, 1995). This sort of strengthening effect could be seen as both paradigmatic and syntagmatic, in that (certain) vowel-intrinsic features are being made more salient, while at the same time the vowel articulation is more clearly set apart from adjacent consonant constrictions, or

other vowels (Beckman et al., 1992; Edwards & Beckman, 1988).

Another well-known form of prominence strengthening on vowels is *hyperarticulation* (Cho, 2005; de Jong, 1991, 1995). A hyperarticulated segment is one in which articulations "[...] enhance the perceptual clarity of the output" (de Jong, 1995, p 493), which will include enhancement of features that are not linked to sonority.[13] In contrasting sonority expansion and hyperarticulation, the following distinction is drawn by de Jong (1995, p 493). Note that "stress" as used by de Jong refers to accentuation at the level of the phrase:

> Stress in the sonority expansion view and in the hyperarticulation view increases phonemic differences. The sonority expansion account chooses certain distinctions having to do with the openness of the vocal tract - "sonority features" - and restricts stress effects to those features. However, unlike the sonority expansion account, the hyperarticulation account predicts that all phonemically distinctive contrasts will be directly affected by stress, not just sonority contrasts.

A clear example of hyperarticulation comes from how non-sonority features are strengthened in prosodically prominent positions. For example, de Jong (1995) tested how American English /ʊ/ was articulated when it bore a pitch accent in a phrase as compared to when it was unaccented (manipulated by a constructed dialogue with corrective focus). It was observed that prominent /ʊ/ generally showed lowering and protrusion of the upper lip, which can be taken as enhancement of the rounding feature of the vowel, not related to sonority. As another example, Cho (2005) finds that American English /i/ shows more fronted lingual articulations when prominent, which could be seen as an enhancement of that vowel's frontness (or [-back] feature in feature terms). These adjustments actually reduce a vowel's sonority, i.e. showing a more closed articulation with increased rounding or tongue tip more closely approaching the roof of the mouth. In both of these cases, it is worth noting that some sonority-expanding gestures are evident as well, where prominent /ʊ/ and /i/ showed

[13]This is conceptually related to the Hyper- & Hypo-articulation (H&H) theory of Lindblom (1990). However, unlike general hyperarticulation, prominence as hyperarticulation is conceived of as *localized*, in that it docks to a single prominent unit. In a language like American English, this unit would be a metrically prominent syllable which is accented in a phrase.

more opening of the jaw and lips respectively (/i/ did not show more more jaw opening under prominence). However, overall we can see clearly that more than just sonority features are being strengthened in these cases. Cho (2005) presents a clear argument that hyperarticulation takes precedence in the case of /i/ at least, in the sense that he found the tongue body was not lowered under prominence, nor was the jaw. Cho frames this as "suppression" of sonority expansion, which could otherwise jeopardize attainment of a high vowel target, and possibly interfere in contrast maintenance with /ɪ/.

As is apparent from the discussion above, the way in which a vowel is strengthened under prominence is dependent on its intrinsic features, where for example prominent articulations of /i/ suppress sonority expansion in terms of tongue and jaw position as found by Cho (2005). This is unlike other vowels which show clear sonority-expanding gestures (Cho, 2005; de Jong, 1991; Erickson, 2002; van Summers, 1987). The relationship between prominence strengthening and a vowel's features is a point that will be returned to below.

As we might expect, articulatory modulations associated with prosodic prominence change the acoustic structure of vowels (e.g., Cho, 2005; Delattre, 1969; Garellek & White, 2015; Mooshammer & Geng, 2008; Nadeu, 2014; van Summers, 1987). For example, sonority-expanding gestures result in a more open oral cavity, raising F1 (van Summers, 1987). Lingual fronting under prominence for /i/ also raised F2 substantially for that vowel (Cho, 2005; Kim, Choi, & Cho, 2016). Just as VOT was described as varying as a function of both prosodic and segmental factors (as shown in Figure 1.1), we could say the same for vowel formants. A vowel's formant structure will vary based on featural specifications (e.g., vowel height, rounding, and so on), and will also vary based on prosodic prominence which systematically fine-tunes how vowel articulations are realized in various ways. This prosodic prominence can be described as phrasal (or, structural) in the sense that the studies mentioned above manipulate contrasts in accentuation (a categorical distinction between accented/unaccented). Nevertheless, as discussed in Section 1.3.1, a bundle of phonetic parameters (including those in the segmental domain, such as VOT) will co-vary with vowel formants to cue this information to listeners. One such phonetic feature that encodes prominence in American English, and also notably varies *within* pitch accent category, is glottal-

ization, discussed below in Section 1.3.4.

1.3.4 Glottalization as prominence strengthening

Another way in which vowels are strengthened under prominence is by changes in voice quality. More generally, various previous studies have argued that voice quality plays an important role in signaling prosodic organization, including boundary marking (Dilley, Shattuck-Hufnagel, & Ostendorf, 1996; Garellek & Seyfarth, 2016; Pierrehumbert & Talkin, 1992; Slifka, 2006), though prominence strengthening is argued here to have more of a connection to glottalization in American English, a point that will be relevant to the experiments in this dissertation. Here a terminological note is pertinent. As discussed in Garellek (2013), a glottal stop can be seen as an articulatory target that can be realized both as a full and sustained stop, as in [ʔ], but may also be realized with incomplete vocal fold closure, manifested as laryngealized voice quality and produced with increased constriction of the vocal folds, as in [ʔ̰]. The term glottalization is used here to refer to the acoustic consequences of a full glottal stop or laryngealized voice quality, following Garellek (2013). A "full glottal stop" will additionally refer to [ʔ].

One well-known pattern is the occurrence of glottalization in vowel-initial words at the beginning of a prosodic domain, and in words that are phrasally prominent (Dilley et al., 1996). In their study, Dilley et al. (1996) used a perceptual criterion for identifying glottalization in a corpus of speech, such that glottalization was coded as being present if there was a full glottal stop, or changes in voice quality, pitch and intensity that led to the percept of glottalization.[14] The authors then observed how the presence/absence of glottalization patterned as a function of phrasal prominence and boundaries. Overall, prominence increased the occurrence of glottalization such that accented syllables, and syllables with full (non-reduced) vowels were more likely to be glottalized. This points to glottalization as a likely manifestation of prominence strengthening. Prosodic boundaries additionally played a role: vowel-initial words at the beginning of higher-level prosodic domains were more likely to be

[14]Both pitch and intensity have been shown to be reliable cues to glottalization in this regard (Gerfen & Baker, 2005; Hillenbrand & Houde, 1996; Pierrehumbert & Frisch, 1997).

glottalized than those at the beginning of lower-level prosodic domains (here described by the authors in terms of break indices based on ToBI annotation; Beckman & Ayers, 1997).

Garellek (2014) notes, however, that the way in which glottalization patterns suggests a tighter relationship to prosodic prominence, as compared to prosodic boundary marking. To measure voice quality, Garellek used electroglottography (EGG), a non-invasive tool for measuring vocal fold contact during voicing. Laryngealized voice quality that would be expected on the basis of glottalization shows increased glottal constriction during voicing and can be indexed by increased vocal fold contact as measured by EGG. In this way EGG allows for a way to quantify glottalization beyond coding its presence/absence, as in Dilley et al. (1996).

In support of the idea that glottalization is linked to prominence, Garellek (2014) found that *less* articulatory contact, measured with EGG, was observed at the beginning of higher phrasal domains as compared to lower phrasal domains in word-initial vowels. This decreased contact, which evidences breathy voicing, is likely attributable to phrase-initial pitch reset, where falling pitch (immediately after pitch reset) results in relaxation of the cricothyroid and thyroarytenoid muscles (Hirano, Ohala, & Vennard, 1969), and vocal fold abduction (Mendelsohn & Zhang, 2011; Zhang, 2011). Given that being phrase-initial overall resulted in *breathier* voicing, Garellek proposes that phrase-initial glottalization serves to counteract the effects of pitch reset on voice quality, explaining its observed prevalence in e.g., Dilley et al. (1996). Breathier voicing leads to decreased intensity and weaker formant energy (Garellek & Keating, 2011; Gordon & Ladefoged, 2001), i.e. decreased perceptual salience (or, phonetic prominence). Glottalization in prominent phrase-initial vowels accordingly "strengthens" voice quality in maintaining more high frequency energy and overall intensity, which enhances formant structure (Garellek, 2011; Hanson, Stevens, Kuo, Chen, & Slifka, 2001). This view of phrase-initial glottalization implicates prominence strengthening as the driving force behind it.[15]

[15]Garellek (2014) notes that it's possible that a glottal stop, without subsequent laryngealization, could be related to prosodic boundary marking, as an increased rate of glottal stops was observed in some non-prominent phrase-initial vowels (Dilley et al., 1996), and it has been observed in German that a glottal stop can be implemented without following laryngealization (Kohler, 1994).

Further in support of this idea, Garellek (2014) found that prominent vowels in general showed increased contact as measured by EGG, as compared to non-prominent vowels in various phrasal positions, evidencing laryngealized voice quality driven by glottalization (analogous to what Dilley et al. found in terms of glottalization and prominence). By comparison, prominent sonorants did not show similar effects. This speaks against voice quality strengthening as simply a byproduct of more forceful articulations (cf. Fougeron, 2001; Fujimura, 1990). If general articulatory effort was driving voice quality changes in prominent word-initial vowels, it should be expected in word-initial voiced segments more generally. Further, given that sonorants undergo various supraglottal modulations in prominence strengthening (Cho & Keating, 2009; Cho et al., 2017), voice quality strengthening may not be "needed" to convey prominence. This further implicates glottalization (as measured by vocal fold contact with EGG) as an intentional form of prominence strengthening (see Garellek, 2013, 2014 for further discussion).

If glottalization represents prominence strengthening on the part of the speaker, we might correspondingly expect that listeners would make use of voice quality as a prominence cue in perception. Studies have investigated the perception and processing of glottalization word-finally (Chong & Garellek, 2018; Garellek, 2015), though the influence of word-initial glottalization in perception, and particularly in the perception of vowel contrasts, has not been explored to my knowledge. Of note, Dilley et al. (1996) show that glottalization does not always occur with accentual prominence, though it tends to. A word-initial vowel can be accented, but not glottalized, and as such glottalization can be seen as a phonetic parameter that patterns with prominence, including *within* phonological (here, accentual) prominence categories. In other words, we can see glottalization as a phonetic prominence cue that might encode granular prominence distinctions (more granular than a binary accented/unaccented contrast). This point is discussed further in Chapter 3. An open question which is tested in this dissertation is accordingly whether word-initial glottalization impacts the perception of subsequent vowels, and if it functions as a cue to prominence in that regard.

1.4 Prosody and segment in perception and processing

Thus far, we've seen the various and detailed ways in which prosodic boundaries and prominence modulate segmental detail in speech. As suggested above, listeners would benefit from distinguishing between the prosodic and segmental influences on a given cue, which might help them parse a segmental message from the speech signal. The claim then, forwarded in various recent studies, though perhaps traceable to its original formulation in Cho et al. (2007), is that listeners would benefit from taking phrasal context into account in segmental processing. That is, given the fact that various phonetic parameters are modulated in a systematic way by prosody, and given that these same parameters cue segmental contrasts, listeners would benefit from reconciling a cue with the prosodic context in which it occurs. Put differently, listeners would benefit from using prosodic information "[...] in determining whether segmental information is driven lexically or post-lexically (prosodic-structurally)" (Mitterer et al., 2019, p 14).

Here it is pertinent to make explicit some assumptions about what is going on during the process of mapping acoustic information in the speech signal to linguistically meaningful units. This framework, outlined below, will help pinpoint the ways in which prosodic context might influence this process.

1.4.1 A general model of perception and processing

One standard assumption in models of speech perception and spoken word recognition is that the processes responsible for the transduction of vibrating air to recognized linguistic units take place in multiple stages. Though the modularity of these various stages and the ways in which information flows between them is a topic of debate (Magnuson, Mirman, Luthra, Strauss, & Harris, 2018; Norris, McQueen, & Cutler, 2016, 2018), several uncontroversial commonalities are reviewed here. Well-known models such as TRACE (McClelland & Elman, 1986) and Merge (Norris, 1999; Norris, McQueen, & Cutler, 2000) share the assumption that the path from signal to understood words involves a mapping from the speech stream to sound categories in a language. These units subsequently activate entries in the mental lexicon.

Simplifying a bit, we could say these models assume listeners first figure out what sounds (or we could say, segments) they are hearing, and then, as those sounds are accumulated, what words are intended by the speaker.

The path from sound waves to linguistic (segmental) categories has gone by various names, e.g., "sublexical phonological abstraction" (McQueen, Cutler, & Norris, 2006), and "segmental parsing" (Rysling, 2017). As these names suggest, the idea is that listeners parse abstract segmental categories from the signal, which in turn feed forward into lexical activation.[16] The literature documents many ways in which this sort of segmental processing, as it will be referred to here, is influenced by context. Consider the fact that contrastive distinctions in segmental categories are signaled by many acoustic cues (e.g., Lisker, 1986; Seyfarth & Garellek, 2018) and that perception of cues is known to be highly *context dependent*. "Context" can be defined broadly to mean information in the speech signal that precedes or follows a given acoustic event of interest (e.g., a cue to a segmental category) in time. Context can span a matter of milliseconds surrounding this point of interest, could be temporally removed from a target, or could even constitute speech material from utterances spoken minutes before (Baese-Berk et al., 2014; Maslowski, Meyer, & Bosker, 2020): all of these sorts of contextual influences have been shown to shape how listeners perceive speech in the process of mapping acoustic cues to a segmental category (see e.g., Stilp, 2020 for an overview).

To illustrate this idea of context-sensitivity, consider a classic example from Mann (1980). In Mann's experiment listeners categorized a stop from a continuum (in a CV sequence) as /dɑ/ or /gɑ/. The continuum varied in only the frequency of the third formant (F3) as a cue to place of articulation, where F3 is lower for /g/. The continuum was preceded by one of two syllables, /ɑl/ or /ɑɹ/. Note that /ɹ/ has much lower F3 than /l/. Mann supposed that a stop coarticulated with these preceding liquids would have changed F3: for example

[16]The idea that listeners are abstracting at this stage runs counter to purely exemplar based/episodic models in which the mental lexicon is only a collection of stored exemplars, e.g., Goldinger (1998); Hawkins (2003). See e.g., Cutler (2010); Cutler, Eisner, McQueen, and Norris (2006); McQueen et al. (2006) for arguments in favor of abstraction. Importantly, the occurrence of abstraction does not mean episodic detail is necessarily discarded (McQueen et al., 2006).

/g/ coarticulated with /l/ would have higher F3 as compared to /g/ coarticulated with /ɹ/. Perceptual compensation for this pattern would entail listeners factoring out the influencing of a preceding liquid when deciding the place of articulation of the following stop (based on F3). Because a /g/ coarticulated with /l/ would have higher F3, listeners would overall accept higher F3 continuum steps as /g/ when /l/ precedes it. These steps (if sufficiently ambiguous) would otherwise be categorized as /d/ when preceded by /ɹ/ (because high F3 following /ɹ/ must cue /d/). This pattern is what Mann found, showing that listeners took contextual F3 into account when perceiving place of articulation. More generally, this shows that listeners reconciled a cue value with its context by attributing the actual F3 value on the continuum to being influenced by preceding material. This is something that is often framed as *compensation* for contextual influences. Context effects of this sort can generally be thought of arising from listeners' learned co-variance of cues that pattern together (Toscano & McMurray, 2010, 2012), or as coming from auditory contrast mechanisms (Holt, Lotto, & Kluender, 2000), which are discussed below.[17]

One common way of modeling this sort of contextual effect is in adjusting a given cue's perceived value, based on the context in which it occurs. In this sense, perception is non-veridical to the extent that a cue's contribution to a perceived segmental category (or featural representation as in e.g., Stevens, 2002) is influenced by context. To make this more concrete, we can consider a model that provides a computationally explicit implementation along these lines. C-CuRE ("computing cues relative to expectations"; Cole, Linebaugh, Munson, & McMurray, 2010; McMurray, Cole, & Munson, 2011; McMurray & Jongman, 2011) is a model that implements compensatory adjustments in the perception of cues wherein a cue value is re-coded based on its deviation from expectations. "Expectations" come from context, where, in model terms, an expected value is derived from by-context regressions (e.g., regressions for a male or female talker, for different consonant contexts, etc.), establishing

[17]An auditory contrast account of Mann's results rests on the finding that frequency distributions in speech are perceived relative to context in a way that is generally *contrastive* (Stilp, 2020). For example, following high F3 in /l/, F3 in the stop will be perceived as relatively low (/g/-like). Following low F3 in /ɹ/, F3 in the stop would be perceived as relatively high (/d/-like). These sorts of effects are discussed in more detail in Chapter 3 and Appendix A.

how a cue is generally realized contextually. A re-coded cue value therefore embodies the listener's knowledge of how context typically influences a cue's realization, and how that given instance of a cue deviates from that norm (in this sense, phonetic detail is retained in that each cue is represented numerically, not as a more abstracted object). C-CuRE was shown to provide the best fit to some speech perception data in comparison to a model created by McMurray and Jongman (2011) that made use of many cues without compensation (similar in spirit to e.g., Hawkins, 2003; Nearey, 1997). This could be taken as evidence that this sort of compensatory adjustment is taking place.

Contexts that elicit compensatory (or contrast) effects of this sort include surrounding speech rate and spectral energy distributions, as in the example from Mann (1980) discussed above (see also Bosker, Reinisch, & Sjerps, 2017; Reinisch & Sjerps, 2013). A classic result showing contextual influences of *speech rate* is that of Miller and Liberman (1979). The authors found that perception of transition duration differentiating /b/ and /w/ (where longer transition duration into a following vowel cues /w/) was modulated by surrounding rate. Slower contextual rate led to a transition duration being perceived as relatively fast/short, cuing /b/. Faster contextual rate led to the transition duration being perceived as relatively slow/long, cuing /w/. Rate-based contextual effects of this sort will be further discussed in Section 1.4.3.

As we might expect from the idea that these effects entail a re-coding of cue values which feed into a perceived segmental category, they are typically described as occurring early in processing, that is prior to sublexical (segmental) abstraction. Empirical evidence for the idea that these acoustic context effects operate early in processing comes from various sources. For example, it has been shown that some acoustic context effects can occur across different speakers' voices (Newman & Sawusch, 1996), before the listener has segregated the speech stream into different talkers, which is assumed to occur early in auditory processing (Bregman, 1994; Cusack, Decks, Aikman, & Carlyon, 2004). Some acoustic context effects are also not impacted by cognitive load (Bosker et al., 2017), and occur with non-speech stimuli (Wade & Holt, 2005), suggesting a general auditory level of processing. Context effects from changing speech rate and spectral context also show rapid influences as measured with

eyetracking (Reinisch & Sjerps, 2013).[18]

We can define context effects of this sort as *pre-lexical*, that is, they involve the processing of cues prior to contact with the lexicon (i.e. in the process of "sublexical phonological abstraction"). We can contrast these effects with other known influences on segmental categorization, those that are *post-lexical*. Recall that segmental categories subsequently activate possible lexical candidates, following e.g., McClelland and Elman (1986); Norris et al. (2000). Here another important influence in word recognition occurs: lexical competition. The notion of lexical competition rests on the assumption that as listeners process speech they entertain multiple possibilities (also called lexical candidates, or lexical hypotheses) for what they are hearing. The consideration of various possibilities that are supported by bottom-up input would help listeners in the face of a noisy and variable acoustic signal (see e.g., Luce & Pisoni, 1998 for an overview). This idea is built into the architecture of many models of perception and word recognition (Luce & Pisoni, 1998; Marslen-Wilson & Tyler, 1980; McClelland & Elman, 1986; Norris, 1994, 1999), and is generally received wisdom in the field. Lexical activation and competition represents another point in speech recognition at which various sources of information can impact processing. Activated word forms "compete" in the sense that they receive varying degrees of support from bottom-up information, from context (broadly construed), and from other factors related to lexical structure, for example word frequency and neighborhood density (Luce & Pisoni, 1998; Vitevitch & Luce, 1998; Vitevitch, Luce, Pisoni, & Auer, 1999).

Consider several examples of effects that are seen as resulting from the lexical competition process. Word that have more phonological neighbors (where a neighbor is a similar sounding word, usually operationalized as a word that is created by removing, adding or substituting one phoneme) are processed more slowly than words with fewer neighbors (Vitevitch & Luce, 1998; Vitevitch et al., 1999). This is argued to arise from increased lexical competition in denser neighborhoods, that is, more word forms under considerations leading to processing slow-downs. Neighborhood density has been argued to exert an influence even in segmental

[18]See also Bosker (2017); Diehl and Walsh (1989); Lehet and Holt (2020); Reinisch (2016) for similar arguments.

categorization. Newman, Sawusch, and Luce (1997) found that in categorizing a VOT continuum that ranged between two sets of non-words (e.g., /kaɪs/-/gaɪs/ and /kaɪp/-/gaɪp/), listeners were biased towards perceiving a non-word with a higher neighborhood density, which the authors viewed as resulting from activated phonological neighbors contributing activation to a denser-neighborhood non-word and biasing categorization. Another type of lexical effect in segmental categorization is word frequency. Connine, Titone, and Wang (1993) found a similar result to Newman and Sawusch (1996) occurs on the basis of frequency. When phonetic continua range between real words, listeners are biased to categorize an ambiguous sound as a higher-frequency word.

These influences on segmental categorization sketched above, and other effects that necessitate access to lexical information such as those related to a word's meaning, have been explicitly argued to operate later in processing (see e.g., Cairns & Hsu, 1980; McClelland & Elman, 1986; Newman & Sawusch, 1996; Swinney, 1979). Given the rough model sketched above, this delay makes good sense: if activation of word forms in the lexicon is required for access to information about neighborhood density, frequency, and so on, we should only see effects such as these occurring once listeners have activated lexical hypotheses and have begun integrating information via lexical competition. Once again, the modularity and information flow between the lexicon and pre-lexical stages of processing is controversial (Magnuson et al., 2018; Norris et al., 2000, 2018), but what is not controversial is the idea that lexical access takes place in multiple stages, and that certain pieces of information can be used by listeners only after contact with the lexicon is made. As it pertains to speech sound categorization, we could define these later influences as *post-lexical* in that they are "decisions about phoneme identity [that] are made on the basis of lexical representations" (Norris et al., 2000, p 320). More generally, contextual effects of this sort could be seen as involving the incorporation of higher-level information (e.g., lexical neighbors), which is brought to bear on the decision process during word recognition.

In describing this very general model of speech recognition, we have pinpointed two stages in which context might influence processing. In one, context effects modulate how acoustic information is perceived, and accordingly how a cue contributes to a perceived segmental

category. In another, contextual information provides support for lexical candidates during lexical competition, on the basis of other higher-level information. With this in mind, we can now turn to an overview of prosodic context effects in speech perception, and the current theory of how listeners integrate prosodic and segmental information.

1.4.2 Prosodic context effects in segmental perception

In light of the assumptions laid out above, we can review some recent findings that show how prosodic context influences speech perception. One influential study is Kim and Cho (2013). In this study, the authors tested how American English speaking listeners' perception of VOT might be sensitive to prosodic context, in particular to the presence or absence of a preceding prosodic boundary. As discussed in Section 1.2.2 (and shown in Figure 1.1), VOT is often lengthened at the beginning of phrasal domains as a function of domain-initial strengthening. Kim and Cho tested whether listeners would accordingly adjust their perception of VOT as a cue to voicing based on whether a target sound was IP-medial or IP-initial. A VOT continuum ranging from American English /b/ to /p/ was categorized as /bɑ/ or /pɑ/ in the carrier phrase "Let's hear x again", where x is the target. Phrasing was manipulated such that, to cue the target as IP-initial, an IP boundary, marked by lengthening and a low boundary tone, was located before the target (such that the carrier phrase contained two IPs). In the IP-medial condition no such boundary preceded the target such that the carrier phrase was made up of a single IP. Kim and Cho found that as predicted, listeners required longer VOT for a voiceless /p/ response when the target was IP-initial, suggesting that they took IP-initial lengthening of VOT into account in their perception of the voicing contrast. This finding points to a possible interaction between prosodic phrasing and segmental perception, that is, that listeners disentangled prosodic and segmental influences on a cue value in deciding how VOT should contribute to perception of the voicing contrast.

Complementing this result, pre-boundary lengthening has also been shown to modulate listeners' perception of durational cues domain-finally. For example, Nooteboom and Doodeman (1980) tested how Dutch listeners' perception of a vowel length contrast shifted as a

function of prosodic position. The authors placed a vowel duration continuum that listeners categorized as phonemically short or long vowel in different prosodic contexts (represented by the authors in terms of syntactic structure). When a target vowel was at the end of a larger phrasal domain, listeners effectively required longer vowel duration to perceive that vowel as phonemically long. Put differently, listeners expected a phrase-final vowel to undergo phrase-final lengthening, lining up with the idea that they disentangled or "factored out" prosodic influences on vowel duration, in similar fashion to the case of VOT and domain initial strengthening in Kim and Cho (2013). Steffman (2019b) further showed that pre-boundary lengthening modulates American English listeners' perception of vowel duration as a cue to coda voicing (where vowels preceding voiced stops are longer and this is a strong cue to voicing for listeners; Chen, 1970; Raphael, 1972). In similar fashion to Nooteboom and Doodeman (1980), longer vowel duration was required for the target to be perceived as voiced when it was phrase final, i.e. expected lengthening for a phrase-final target sound led to longer required vowel duration to perceive voicing. Steffman and Katsuda (2020) further showed similar effects for Tokyo Japanese speakers' perception of a vowel length contrast, where prosodic phrasing was cued only by changes in contextual pitch.

The current literature therefore offers general empirical support for the notion that phrasal prosody fine-tunes the perception of segmental contrasts, in correspondence with the way it fine-tunes segmental articulations in speech production. As mentioned above, however, these findings have only focused on the influence of prosodic boundaries.

1.4.3 Considering domain-general effects

In light of the results outlined above, one relevant consideration is the idea that prosodic context might engender perceptual shifts by a mechanism that does not involve listeners' direct reference to prosodic structure. This point is raised in detail by Mitterer, Cho, and Kim (2016), who note a possible alternative explanation for the effect of phrasing on VOT perception found in Kim and Cho (2013). Recall that in Kim and Cho's IP-initial condition, the target was preceded by phrase final lengthening (cuing the fact that the target was

phrase-initial): a longer segment therefore preceded the target in the IP-initial condition as compared to the IP-medial condition. This difference in pre-target duration offers a possible explanation for the effect observed by Kim and Cho via the mechanism of *durational contrast* (e.g., Bosker, 2017; Diehl & Walsh, 1989). This refers to a perceptual effect whereby, generally speaking, the perceived duration of an acoustic event is affected contrastively by the duration of surrounding material (as in Miller & Liberman, 1979, described in Section 1.4.1 above). For example, a longer vowel preceding the target segment in Kim and Cho's stimuli would make VOT be perceived as relatively short, as compared to a shorter preceding vowel (e.g., Steffman, 2019a; Toscano & McMurray, 2015). This would predict shorter perceived VOT in the IP-initial condition, leading to decreased aspirated responses therein. As such, the observed effect could be seen as an indirect consequence of temporal patterns associated with prosodic boundaries, and importantly would not result from listeners' reference to prosodic structure.[19] Mitterer et al. (2016) accordingly frame this as a case that is ambiguous in terms of what processes are responsible for the categorization data: on one hand, listeners could be referencing higher-order prosodic structure, and on the other, domain-general durational contrast effects could be influencing perception of VOT. Given that the present dissertation is concerned with how listeners' perception of spectral information is impacted by prosodic context, another domain-general contrast effect, *spectral contrast* (e.g., Holt, 2006) may additionally be relevant. Spectral contrast refers to contrast effects in which frequency distributions in the spectrum are perceived relative to their context. This will be discussed in more detail in Chapter 3.

Mitterer et al. (2016) more generally raise the thorny issue of teasing apart other influences in segmental perception from influences that involve listeners' reference to prosodic structure, or a prosodic property like prominence (see Mitterer et al., 2016; Steffman, 2019a; Steffman & Jun, 2019 for further discussion of this point). This is particularly relevant in cases where the property in question involves any temporal changes, something that can be hard to avoid given the tight relationship between prosody and temporal patterns in speech

[19]Durational contrast effects are assumed to arise early in general auditory processing by virtue of the fact that they occur across speakers before auditory stream segregation (Diehl, Souther, & Convis, 1980; Newman & Sawusch, 2009), and with non-speech stimuli (Bosker, 2017; Diehl et al., 1980).

(Cho, 2015; Fletcher, 2010). We can consider several avenues for addressing this issue. One route taken by several recent studies is to manipulate pitch only as a cue to prosodic context (Kim, Mitterer, & Cho, 2018; Steffman, 2019a; Steffman & Katsuda, 2020), given that pitch patterns play an important role in delimiting prosodic boundaries, as described in models of languages' intonational phonology (Beckman & Pierrehumbert, 1986; Jun, 1996, 2005, 2014; Pierrehumbert, 1980). By holding duration constant across conditions that manipulate prosodic context, the influence of durational contrast is controlled, though possible influences of changing pitch on perceived duration should be considered, given that differences in pitch alone can influence the perception of duration in some cases (cf. Brigner, 1988; Steffman & Jun, 2019; van Dommelen, 1993; see Steffman, 2019a for discussion). Another option is to construct stimuli in such a way that contrast effects predict the opposite of what would be expected based on prosodic context. If e.g., a durational contrast effect predicts the opposite of what is observed, then it can clearly be ruled out as a possible explanation. This strategy was used by Steffman (2019b), and in several experiments which follow in this dissertation. Another possible way to disentangle these effects is to look at online processing measures (e.g., eye movements in an eyetracking study), as done by Kim, Mitterer, and Cho (2018) and Mitterer et al. (2019). Because contrast effects are known to operate early in processing (Bosker et al., 2017; Reinisch & Sjerps, 2013), deviations from this timecourse might suggest other processes are responsible for an observed effect. This point is discussed in detail below.

In spite of the possible influence of durational contrast in Kim and Cho (2013), the fact that predicted prosodic context effects did indeed emerge in all of the studies mentioned above suggests that listeners really are making reference (in some way) to prosody in segmental processing, though the idea that prosodic structure exerts a direct influence is one which merits further empirical support. This is one basic goal of the current dissertation.

In summary, the possible influence of other context effects that result from the manipulation of temporal and spectral context merit consideration in studies such as those contained in this dissertation. In cases where the goal is to rule them out as a possible influence, disentangling them in some way from the predicted effects of prosodic context will be important. This point is returned to throughout the dissertation.

1.4.4 How are prosody and segment integrated?

Given the literature reviewed above, the apparent involvement of phrasal prosodic structure in segmental processing raises various questions about how listeners are building and integrating both prosodic and segmental representations as they listen to speech. The findings outlined in Section 1.4.3 suggest that listeners integrate segmental and prosodic structure in some way. However, this leaves open the question of how both prosodic structure and segmental categories are being processed.

Cho et al. (2007) propose a mechanism, the *Prosody Analyzer*, to explain the influence of prosodic structure in listeners' word segmentation and lexical access. Their motivating experimental evidence is a cross-modal priming experiment. The authors tested how an IP boundary (versus word boundary) influences word segmentation, and found that listeners more quickly recognize a word, in a temporarily ambiguous two-word sequence e.g., "bus" from "bus # tickets", (which is temporarily ambiguous with "bust"), when initial /t/ manifests IP-initial strengthening patterns. This finding suggests that the acoustic consequences of a larger prosodic boundary facilitate word segmentation. The authors propose that listeners extract parallel prosodic and segmental representations, using whatever information they have available in the speech signal to specify both types of representations (IP-initial VOT would be useful for specifying both the prosodic boundary and segmental representation in this case).[20]

The prosodic analysis model (as it will be called here) proposes that integration of prosodic and segmental representations occurs via the process of lexical competition. The segmental representation that is parsed out of the signal presents possible candidates, e.g., "bus" and "bust", and is combined with a prosodic representation, which encodes an IP boundary between /s/ and /t/. The Prosody Analyzer therefore indicates where words are likely to begin and end, as a function of prosodic boundary, and this information is used to help decide between possible words, facilitating word segmentation. The role of prosody

[20]The idea that these structures are processed "in parallel" can be taken to mean that listeners construct a representation of each structure simultaneously as speech unfolds, as described in Cho et al. (2007). See also Christophe et al. (2004); Kim, Mitterer, and Cho (2018); Salverda et al. (2007).

more generally in this model is to help select between possible word candidates, as a function of the context in which they occur. In this sense we can conceptualize the role of prosody as entering relatively late in the word recognition process, i.e. as helping listeners to decide between lexical candidates which are activated on the basis of segmental information. The model therefore demarcates the role of prosodic and segmental representations as entering at different stages in the process of word recognition (though note again that this is agnostic as to what listeners use as cues for the purpose of specifying each type of representation). As stated by Mitterer et al. (2019, p 14):

> [...] although the segmental and prosodic analyses may take place in parallel, their effects do not seem to come into play simultaneously: the segmental analysis activates all possible lexical hypotheses, and its activation is further modulated by the prosodic analysis at a relatively late stage in spoken-word recognition.

The idea that segmental information contributes to activation of lexical candidates, while phrasal prosodic information enters later in the process of spoken word recognition, is consistent with other findings showing that word recognition integrates various sources of information in multiple stages as discussed above. In the prosodic analysis model, a parsed-out prosodic structure would constitute a similar modulating influence, which may also be used by listeners in other domains of processing, e.g., in syntactic processing (see e.g., Schafer, 1997; Speer, Kjelgaard, & Dobroth, 1996; Speer, Warren, & Schafer, 2003, and discussion in Cho et al., 2007). Two recent studies, described below, present time-course evidence in support of this sort of later-stage prosodic analysis in segmental processing.

Kim, Mitterer, and Cho (2018) tested how Korean listeners use tonal prosodic cues in a phonological inferencing task. They tested Korean post-obstruent tensing (POT), whereby lax (also called lenis, as in Figure 1.1) stops and affricates become tense (also called fortis) following another obstruent. For example, /puri/ "beak" will become tensified [p*uri] when following an obstruent, as in the sequence /porasɛk puri/ "purple beak", making it more or less homophonous with /p*uri/ "root". The domain of this process is the accentual phrase (AP, see Jun, 1996, 1998). Kim et al. implemented a visual world eyetracking exper-

iment in which listeners heard a tensified obstruent (underlyingly lax) and were presented visually with both an orthographic tense form and an orthographic lax form. The authors were curious if listeners would look to the underlying representation (effectively undoing the application of POT), or would look to the surface (tense) representation. Crucially, the authors predicted that this process should be modulated by prosody: because the domain of POT is the AP, listeners would necessarily need to reference accentual phrasing to determine whether POT applied. For example, AP-internal [p*uri] as in (porasɛk p*uri) where parentheses indicate an AP boundary, may be "beak" or "root". However, an AP-initial target word disambiguates the meaning: (porasɛk) (p*uri) can only be "root". Kim et al. found that when listeners heard a phonetically tense target [p*uri], they looked more to an underlyingly lax word /puri/ when that word was in an AP-internal context that licensed POT (e.g., (porasɛk puri)), as compared to when it was not. This evidenced the predicted phonological inferencing effect. In another experiment, the authors showed that this effect goes away when an AP boundary intervenes between the potential target for POT and a preceding obstruent. This may be taken to suggest that the observed phonological inferencing effect makes reference to the phrasal domain of the AP. The authors find that this prosodically modulated inferencing effect occurs relatively late in processing, reaching significance approximately 800 ms after target onset. This delayed effect in prosodically-modulated phonological inferencing supports the idea that, as stated by Kim et al., "[...] prosodic structure is parsed in parallel to the segmental level and is used later for prosodic modulation in lexical access" (p 26).

In another study testing the timecourse of prosodic influences in segmental processing, Mitterer et al. (2019) explored how listeners use prosodic boundaries in Maltese. In this language a glottal stop can signal phonemic contrasts (e.g., /ɑːm/ "he swam" versus /ʔɑːm/ "he woke up"), while also occurring in vowel-initial words when they are phrase-initial, as a form of initial strengthening (cf. Dilley et al., 1996; Pierrehumbert & Talkin, 1992). In one experiment, Mitterer et al. found that listeners used preceding boundary information in determining if glottalization cued a segmental contrast (in similar fashion to VOT as described above). That is, if a glottalized vowel-initial word was phrase-initial, listeners

attributed glottalization to being phrasal in nature (not cuing a lexical contrast), and thus categorized the word as, e.g., /ɑːm/. When a glottalized vowel-initial word was phrase-medial, listeners interpreted glottalization as contrastive (as it would not be realized as a function of prosodic structure phrase-medially), and categorized the word as, e.g., /ʔɑːm/.

Notably, these effects disappeared in another experiment, a visual world eyetracking task, when following material disambiguated the target word. For example, when listeners have heard only [ʔɑ...] the target word could be either /ʔɑbɑd/ or /ɑbɑʔ/, with the latter being glottalized initially due to the presence of a prosodic boundary. However when the final consonant [d] is heard the word will unambiguously be /ʔɑbɑd/ (because /ɑbɑd/ is not a word). The authors supposed this lack of an online effect may have been because the items they used became lexically disambiguated too early to show an effect of prosodic boundary. In other words, listeners heard material that disambiguated the target word (e.g., the final [d] in the example above) at a point that allowed them to make a lexical decision before the effect of prosodic structure was evident (consistent with the view that prosodic boundary computation should show a later effect in processing). The authors tested this by using the same items in a gating task with disambiguating material masked by noise. Listeners had to guess which word the speaker intended without disambiguating segmental information. Here, the expected effect of prosodic boundary was observed. Listeners were more likely to perceive the target as containing an underlying/contrastive glottal stop when boundary cues were absent (i.e. when a glottal stop would not serve as a phrasal boundary marker). Given that this effect emerged with the same stimuli used in the eyetracking experiment, this offers further support to the idea that prosodic boundary effects occur at a later stage in the word recognition process, too late to be observed online with Mitterer et al.'s materials. Consistent with the proposed function of the Prosody Analyzer, Mitterer et al. note these results support a model in which lexical access takes place in multiple stages and incorporates multiple types of information, with prosodic information being used at a later stage.

This, taken together with Kim, Mitterer, and Cho (2018), offers clear support for a later-stage influence of prosodic information in speech processing, supporting the model proposed by Cho et al. (2007). These findings are more generally consistent with the notion that

higher-level processing (of various kinds) should show a delayed timecourse in its influence, as argued for various other context effects (Bosker et al., 2017; Green, Tomiak, & Kuhl, 1997; Maslowski et al., 2020).

1.4.5 Should prominence and boundary processing be different?

The current evidence thus favors an account in which prosodic structure enters into processing at a relatively late stage, following the activation of lexical candidates on the basis of segmental information, though, as noted above, previous studies focus on the boundary-marking function of prosodic structure. The prominence-marking function of prosodic structure remains unexplored: should we expect prosodic prominence to behave differently?

If we consider prosody as an organizational structure that listeners parse out of the signal (Beckman, 1996; Cho et al., 2007), then the prominence relations encoded in such an abstract representation can be presumed to be analogous to the computation of prosodic boundaries, or part of the same computed structure. With this view alone, we might predict the same sort of perceptual processing as evidenced in Kim, Mitterer, and Cho (2018) and Mitterer et al. (2019) for prominence structure, namely a delayed influence of prominence information reflecting integration of prosodic structure in lexical competition.

However as discussed in Section 1.3.1, prominence perception varies in a continuous and multidimensional fashion. We should therefore not assume that *only* abstract prosodic structure is conveying prominence information to listeners. Prominence at a phonetic level accordingly merits consideration as a possible factor in listeners' perception of segmental contrasts in speech. Indeed, Steffman and Jun (2019) found that pitch height in isolated words (divorced from a phrasal context which would signal phonological prominence structure) shaped how listeners perceived vowel duration as a cue to coda stop voicing in American English. With prominence-lending high pitch on a vowel, listeners expected a longer vowel duration (i.e. as a co-occurrent prominence property), and this shifted their perception of the voicing contrast. This suggests that phonetic prominence information (as compared to phonological organization in a phrase) can shape listeners' perception of segmental cues.

As discussed in Section 1.3, prominence is relational. That is, cues should be perceived as phonetically prominent, or not, depending on their relation to the context around them. For example, the finding that pitch perception is relative to pitch range and context is well established in the psychoacoustic literature (Plantinga & Trainor, 2005; Repp, 1997; Schellenberg & Trehub, 2003), and the context-dependence of duration perception is also well-established (Bosker et al., 2017; Diehl & Walsh, 1989; Jones & McAuley, 2005). Even in the case of Steffman and Jun (2019), where listeners heard a single isolated word in a given trial, perception of prominence in this word would be relative to stimuli heard on other trials, i.e. the global context of the experiment (cf. Bigand & Pineau, 1997; Jones & McAuley, 2005). We could accordingly conceptualize phonetic prominence perception as entailing the relation of a given cue or linguistic unit to its context, e.g., as comparing relative differences in duration and pitch (Diehl & Walsh, 1989; Repp, 1997). We see further evidence for the relative nature of prominence perception from Mo (2011), who found that a measure of syntagmatic prominence relative to adjacent spans of speech best accounted for perceived prominence in an RPT task.

As such, phonetic prominence perception may not involve reference to an abstract (phonological) prosodic structure, presenting a possible difference from boundary processing, though as outlined above, this is an open question. How might listeners integrate phonetic prominence information with their perception of segmental cues? If phonetic prominence perception is considered a more generic acoustic context effect, we might expect to see this sort of prominence information integrated rapidly. As an illustrative example of how acoustic context effects influence online processing, we can consider Toscano and McMurray (2015). Note that this study does not deal with prosody, but nevertheless presents a useful time-course comparison. The authors tested how listeners' perception of the American English /p/-/b/ contrast was influenced by VOT along a continuum, preceding speech rate, and the duration of the following vowel. The authors observed at what point in time these pieces of information became useful to listeners, using a visual world eyetracking paradigm in which listeners looked to images. Here it's pertinent to make explicit a linking hypothesis: in a visual world eyetracking study, increased looks to a target word (whether an image or or-

thographic representation) are taken to index lexical activation. The point in time at which listeners start to look to a given word is therefore taken to be the point in time at which information in the signal becomes useful to them for recognizing that word. Also important in this view is the assumption that lexical activation is not all or none, but can consist of partial decisions, indexed by an increase in looks to a target.

Recall that speech rate impacts perception of durational cues, such as VOT (discussed in Section 1.4.3). The authors accordingly found an effect of preceding speech rate on listeners' perception of the /p/-/b/ contrast. The point in time that it influenced looks to the target was the *same* as the point in time as VOT itself impacted listeners' looks. Even though listeners received preceding speech rate information before they heard VOT, speech rate itself did not contribute to lexical activation, as would be expected given that rate itself should not inform about following lexical material.[21] Simultaneous effects of speech rate and VOT were taken by the authors to reflect the modulation of VOT perception on the basis of preceding rate, i.e. a re-coding of VOT cue values via expectations generated from speech rate. More generally then, the influence of such a contextual factor should therefore be simultaneous with the cue that it modulates.

Though Toscano and McMurray (2015) do not test contextual factors related to phrasal prosody or prominence, context conveying relative (phonetic) prominence information might constitute an analogous influence on listeners' perception of segmental cues, following the assumption that perception of prominence is not solely dependent on a computed prosodic structure. The crucial difference between this account and a prosodic analysis account would be the point in time at which prosodic information is used by listeners. According to prosodic analysis, the influence of phrasal prosody follows lexical activation. Or put differently, prosodic context effects are *post-lexical*. On the other hand, if prominence directly modulates (or re-codes, following C-CuRE) cue values, it should show a relatively early influence in perception, in tandem with segmental cues (analogous to the effect of speech

[21]Note this pattern is different than the one which the authors observe for the duration of the following vowel, which the timecourse evidence suggests is integrated with VOT as an independent cue to the voicing contrast (see also Toscano & McMurray, 2012).

rate and VOT in Toscano & McMurray, 2015). An influence of contextual prominence that was simultaneous with that of an intrinsic segmental cue could in that sense be defined as *pre-lexical*, that is, directly shaping how a cue is perceived.

In summary, given the complex and multidimensional nature of prominence, which not only reflects phrasal (phonological) organization but also derives from acoustic/phonetic context (among other things), it is an open question if prosodic prominence will pattern like prosodic boundaries in showing a delayed influence in processing, or will show an immediate influence, following other acoustic context effects. By investigating the timecourse of prominence effects on word recognition, we can inform a model of its integration in processing, and more specifically, test if its influence is pre- or post-lexical.

Notably, phonological and phonetic prominence often go hand-in-hand: a phonologically (or, structurally) prominent unit will be signaled as such by a bundle of phonetic cues (though here again, phonetic properties can also vary *within* a phonological prominence category). In testing a case where both phonological and phonetic prominence co-vary, we can use timecourse data to address how prominence is being represented and processed. In other words, do listeners show processing consistent with immediate/phonetic effects, or with abstract prominence information accessed through prosodic analysis? Further, in comparing a co-varying case such as this to a "purely" phonetic prominence cue, that is, phonetic variation within a phonological category, we can further test the extent to which prominence processing, as it pertains to segmental perception, is reliant on phonetic information. More precise time-course predictions are discussed in Chapter 2.

Prosodic prominence, being at "the crossroads of signal and structure" as stated by Baumann and Cangemi (2020), thus offers an interesting test case which will extend recent studies that have focused on the influence of prosodic boundaries. In testing how contextual prominence factors into segmental processing, we can hope to enrich the current theory of prosodic analysis and to address how listeners are representing prominence in this domain.

1.5 Goals and scope of the dissertation

There are many ways in which the possible inter-relatedness of prominence and segmental information in processing could be tested. This dissertation limits itself to the question of how speech sound categorization is mediated by prominence. As is apparent from the model of speech recognition sketched in Section 1.4.1, one key building block of spoken language comprehension is deciding what segmental categories are intended by a speaker (see e.g., Cutler, 2010; Rysling, 2017). Decisions about segmental material are influenced by both pre-lexical and post-lexical effects, as described above. As such, this presents one fruitful way of testing if and how prominence information is incorporated in spoken language processing. This dissertation thus tests throughout if, and when in processing, listeners make decisions about segmental categories on the basis of various types of prominence information.

This dissertation restricts itself to investigating prominence that is *contextual*, that is, not temporally co-occurrent with a given to-be-categorized speech sound. This contrasts with a possible approach exploring co-occurrent prominence cues (e.g., pitch and duration of a given segment, as in Steffman & Jun, 2019), though future work could draw on these results to explore how both co-occurrent and contextual prominence cues are processed by listeners. As described above, the dissertation also adopts vowel contrasts as a test case, examining how listeners' perception of vowel categories shifts in the basis of contextual prominence, which will inform a more general theory of prominence in segmental processing, and the architecture for a model of prominence integration.

To frame the goals of the dissertation in a more specific fashion, consider the questions below.

(1) *Does prosodic prominence mediate perception of vowel contrasts?*

This is the general empirical question that experiments throughout this dissertation will answer. We have seen above how prosodic boundaries mediate perception of durational contrasts, and influence processing more generally (Kim & Cho, 2013; Kim, Mitterer, & Cho, 2018; Mitterer et al., 2019; Steffman, 2019b). However it is currently unknown

whether prominence information might similarly factor into speech perception and processing. The basic question that will be tested throughout this dissertation is accordingly if the prominence of a given target sound, in relation to its context, shifts listeners' perception of that sound. If the answer to this question is yes, we will have new evidence for the ways in which prosodic features are incorporated into segmental perception. The perception of spectral cues as influenced by prosodic context also remains untested such that the present experiments are an extension of past studies in this regard.

(2) *How is prominence integrated with segmental cues?*

Or put differently, are prominence effects pre-lexical or post-lexical? This question will be addressed by tracking listeners' eye movements in a visual world eyetracking paradigm. If we see that prominence exerts only a later-stage influence in perception, in line with prosodic analysis as described in Cho et al. (2007), we would have evidence that prominence-lending prosodic context is processed as an abstract phonological structure. Conversely, if we see prominence shows the same timecourse and trajectory as vowel-intrinsic formant cues, we would have evidence that its influence does not include more abstract (post-lexical) integration of prosodic structure. Instead, we could conclude that listeners process prominence as a more general phonetic context effect, without parsing it as part of abstract phonological organization (at least in the domain of segmental processing). This question is addressed in Chapter 2.

(3) *Does segmental context (glottalization) cue prominence?*

As is evident from the discussion in Section 1.3, various manifestations of prominence strengthening might be expected to cue prominence to listeners, with one promising test case being glottalized voice quality and/or [ʔ] (Dilley et al., 1996; Garellek, 2013, 2014). This could be construed as a localized (or, segmental) prominence cue, that might influence vowel perception via its patterning with prominence. Testing whether glottalization cues prominence in a similar fashion to phrasal/phonological organization will help us better understand exactly what sorts of prominence information listeners

are integrating in their perception of vowel contrasts. This question will be addressed in Chapter 3.

(4) *Does prominence processing vary based on prominence-lending context?*

Given the multidimensional nature of prominence perception discussed above, we can ask if localized cues to prominence are processed in the same fashion as more global/phonological prominence-lending context. Following the model of prosodic analysis sketched above, we might expect that an abstract prosodic structure should modulate post-lexical processing as described in Cho et al. (2007) and Mitterer et al. (2019). However, localized cues to prominence such as glottalization might be expected not to factor solely into post-lexical processing but to be incorporated rapidly with segmental cues. This is an entirely open question, which is tested Chapters 2 and 3.

(5) *Do perceptual prominence effects vary based on vowel-intrinsic features?*

As outlined in Section 1.3.3, vowels are strengthened by prominence in various ways depending on their features, and their relationship to the vowel system of a language. This shows the importance of considering vowel-intrinsic featural specifications in describing how prominence strengthening operates (Cho, 2005; de Jong, 1995). Accordingly, one core part of understanding the perceptual effects of prominence strengthening will be explaining how vowel-intrinsic featural properties mediate listeners' perception of prominence and prominence strengthening. Put differently, do listeners apply the same perceptual processing to vowels regardless of vowel features, or do prominence effects take into account the relationship between vowel features and prominence strengthening that we see playing out in the speech production literature? These questions are addressed in Chapter 4.

As these questions are answered, we will consider their implications for the relationship between prosody and segment in perception, the listener's task of integrating the two, and desiderata for a model of perception and processing that accounts for the data. The answers to these questions will further inform a proposal describing listeners' integration of prominence in segmental processing, which will be outlined in Chapter 5.

CHAPTER 2

Phrasal prominence effects online and offline

2.1 The experiments in this chapter

The goal of the two experiments presented in this chapter is to test how phrase-level prominence mediates listeners' perception of segmental information, in particular formant cues to a vowel contrast. As described in Section 1.3.3, vowel articulations, and their acoustic consequences, are modulated in a systematic way by phrasal prominence. This could be seen as analogous to boundary-driven modulation of VOT (in English and Korean) or glottalization (in Maltese): it represents a case in which a cue that is used to make segmental contrasts varies in a systematic way based on prosodic factors. One basic question addressed in this chapter is accordingly if listeners are sensitive to this pattern in perception of vowel contrasts.[1] If the answer is yes, this would provide a new piece of evidence showing the involvement of phrasal prominence in segmental perception. This would further extend past findings, which primarily test durational cues, to examine how the perception of spectral properties is shaped by prosody.

The second core question addressed here relates to how phrasal prominence information is processed by listeners. The timecourse of listeners' use of contextual prosodic information will be assessed in comparison to their use of vowel-intrinsic formant cues, with the goal of testing how these various sources of information influence processing over time, following the two possibilities sketched in Chapter 1. This will be discussed in Section 2.3.

[1]Though notably, previous studies have shown that prosodic prominence in vowels enhances speech intelligibility (e.g., Connaghan & Patel, 2017), in line with the general view of prominence strengthening as facilitating recognition of important/informative parts of the speech stream (e.g., Cutler, 1976; Ladd, 2008).

2.1.1 The test case: Sonority expansion in vowel articulations

As discussed in Section 1.3.3, prominence at the level of the phrase has various effects on vowels, reviewed briefly here. These effects are generally seen as serving the purpose of *paradigmatic* contrast enhancement, though *syntagmatic* enhancement effects may also arise as a function of prominence strengthening (Cho, 2005; de Jong, 1991, 1995; de Jong et al., 1993; Roessig, Mücke, & Pagel, 2019). Recall that syntagmatic enhancement effects are those that help a given segment stand out in relation to surrounding context. Paradigmatic effects are those that help a segment strengthen acoustic properties that are relevant in featural contrasts: for example increased lip-rounding on a vowel like /ʊ/, enhancing its contrast with un-rounded vowels in a given language. In feature terms, this would be seen as enhancement of [+round] (de Jong, 1991, 1995). As noted in Section 1.3.3, modulations that a vowel articulation undergoes when phrasally prominent are dependent on properties of the vowel itself, and its relation to other contrasts in the language (Cho, 2005; de Jong, 1995; Fougeron & Keating, 1997; Garellek & White, 2015).

One well-documented pattern in the literature that is often framed as resulting in syntagmatic contrast enhancement is *sonority expansion* (Cho, 2005; de Jong et al., 1993), where sonority is defined in articulatory terms, following Silverman and Pierrehumbert (1990). Recall that expanding sonority in a vowel articulation entails increased amplitude of jaw lowering, and lowering and backing of lingual articulations in the mouth, allowing more energy to escape (Cho, 2005; de Jong et al., 1993; Erickson, 2002; van Summers, 1987). This can be seen as enhancing syntagmatic contrasts with both adjacent consonant articulations, and other non-prominent vowels.[2] Typically, non-high vowels, when phrasally prominent (i.e. bearing the nuclear accent in a phrase), show sonority expansion as compared to vowels that are unaccented (Erickson, 2002; van Summers, 1987).[3]

[2]Sonority expansion could also be framed as selectively enhancing the "sonority features" of a vowel, that is, enhancement of "distinctions having to do with the openness of the vocal tract" (de Jong, 1995, p 493).

[3]As noted in Chapter 1, this pattern does not necessarily occur for high vowels, where sonority expansion might jeopardize attainment of the articulatory target for the vowel gesture: in these cases sonority expansion can be suppressed (Cho, 2005), or other prominence enhancement effects, e.g., hyperarticulation, are observed (de Jong, 1991, 1995). These effects will be examined in Chapter 4.

One acoustic consequence of sonority expansion is accordingly a change in vowel formant structure. Jaw lowering and lingual backing and lowering correlate with raised first formant (F1) frequencies and lowered second formant (F2) frequencies, and indeed, prominence has been shown to alter the formant structure of vowels in this way (Cho, 2005; Lehiste, 1970; van Summers, 1987). An additional source of perceptual evidence for these effects comes from Mo, Cole, and Hasegawa-Johnson (2009) who tested how changes in formant structure influenced P-scores in an RPT task (indexing perceived prominence as discussed in Section 1.3.1). They observed that, within a vowel category, F1 raising and F2 lowering correlated with an increase in perceived prominence, for (non-high) vowels which undergo sonority expansion. Listeners' perception of prominence in vowels therefore seems to incorporate F1 and F2, and line up with how formant structure is modulated by phrasal prominence in speech production.

Given the influence of phrasal prominence, via sonority expansion, in modulating vowel formants, we could conceptualize F1 and F2 as varying both on the basis of a prosodic dimension (prominence) and a segmental dimension (contrastive vowel categories). This is analogous to the case of VOT shown in Figure 1.1, where a given VOT value is determined not only by segmental category, but also prosodic configuration. In light of this, we can test how changes in a vowel's prominence in a phrase shift listeners' perception of F1 and F2. In other words, we can explore if listeners take into account how sonority expansion, in prominent contexts, has shaped a vowel's realization (for vowels which undergo sonority expansion). Experiment 1 accordingly tests if listeners incorporate phrasal prominence in their perception of formants in segmental processing.

2.2 Experiment 1

2.2.1 Materials

The materials used in Experiment 1 were created by re-synthesizing the speech of a ToBI-trained American English speaker. The speech material was recorded in a sound-attenuated

booth in the UCLA Phonetics Lab, using an SM10A Shure^{TM} microphone and headset. Recordings were digitized at 32 bits and a 44.1 kHz sampling rate.

The vowel categories chosen as a test case are American English /ɛ/ and /æ/. Generally speaking, /æ/ has higher F1 and lower F2 relative to /ɛ/ (Hagiwara, 1997; Peterson & Barney, 1952; Yao, Tilsen, Sprouse, & Johnson, 2010), i.e. it is a lower and less-front vowel.[4] In Experiment 1, listeners' task was to categorize a sound drawn from a continuum as "ebb" /ɛ/ or "ab" /æ/.[5] The continuum for the target word was created by re-synthesizing the formant values of natural speech, such that one endpoint had F1 and F2 which were matched to a naturally produced /ɛ/, spoken by the speaker who produced the model utterances for the materials. The other endpoint had F1 and F2 which were matched to a naturally produced /æ/. The continuum varied jointly in F1 and F2 between each endpoint in 8 interpolated steps (for 10 steps total including endpoints). Each target word was originally recorded in two carrier phrases. These are shown with ToBI labels (Beckman & Ayers, 1997) in (1) and (2) below, where x represents the target word.

(1) I'll say x now
H* H* L-L%

(2) I'll SAY x now
L+H* L-L%

Two phrasal prominence conditions were created in Experiment 1, corresponding to (1) and (2) above. In (1), the target bears relative prominence, being in the nuclear pitch accented (NPA) position of the phrase, which contains a standard declarative tune. In (2), the target follows narrow focus marking, realized with a rising L+H* accent on the word "say"; the target is therefore post-focus, unaccented, and non-prominent. These two conditions,

[4]It can be noted that there is clearly regional variation in terms of how this contrast is manifested in F1 and F2 (Clopper, Pisoni, & de Jong, 2005; Hagiwara, 1997), and duration also plays a role in the contrast where /æ/ is longer (e.g., House, 1961; Umeda, 1975), a point that will be discussed below. Nevertheless, the distinction between these vowels in terms of F1 and F2 appears to be robust, and it will accordingly be assumed that listeners use F1 and F2 to distinguish these vowel categories, with higher F1 and lower F2 signaling /æ/.

[5]These two words were chosen to be relatively matched in frequency, as calculated from the $\text{SUBTLEX}_{\text{US}}$ corpus (Brysbaert & New, 2009). The $\log_{10}$ frequency of "ebb" is 1.28, and the $\log_{10}$ frequency of "ab" is 1.88 (both words are quite low frequency). Importantly, any frequency bias would be expected to impact overall responses, but not to mediate the effect of prominence.

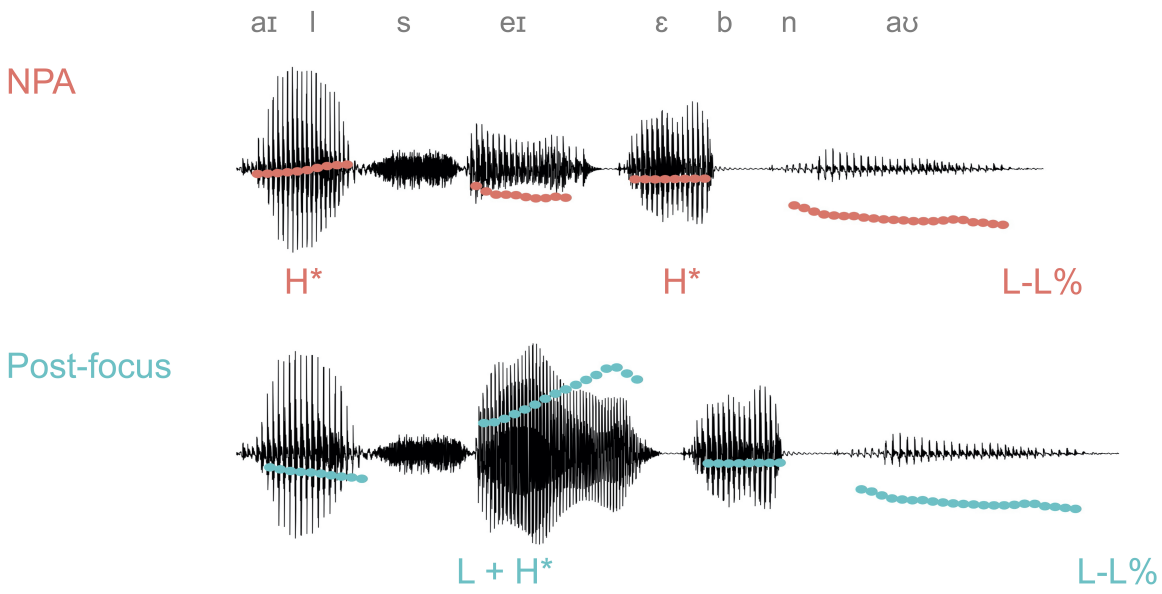

Figure 2.1: Waveforms of the Experiment 1 stimuli, overlaid with pitch tracks in both conditions. The "nuclear pitch accent" (NPA) condition, in which the target is prominent, is shown at the top. The post-focus condition, in which the target is non-prominent, is shown at the bottom. A segmental transcription is given in IPA above, aligned to the top-most waveform. The pitch range spans the maximum amplitude in the waveforms and is 50 to 250 Hz. The target word shown in the figure is from the /ɛ/ endpoint of the continuum.

referred to as the NPA and post-focus condition, were created by cross-splicing and PSOLA method resynthesis in Praat (Boersma & Weenink, 2020; Moulines & Charpentier, 1990).

The goal in creating these conditions was to manipulate only the context surrounding the target (with the target identical across conditions), in such a way that listeners' perception of target prominence was roughly equivalent to the (phonological) ToBI-labeled examples in (1) and (2). These stimuli accordingly present a fairly conservative manipulation, changing only context to ensure that properties of the target sound itself did not shift listeners' perception. Any differences observed across conditions in the experiments that follow can only be attributed to context.

Two different frames were created, corresponding to (1) and (2), where "frame" refers

to the carrier sentence surrounding the target word. The starting point for the creation of these frames was (1) above. The NPA condition was created simply by using the frame in (1), from which the target sound was excised. To create the post-focus condition, the vowel in "say" from (2), with narrow focus, was spliced into the frame, replacing the vowel in "say" from (1). The vowel in "say" in the post-focus condition therefore has increased amplitude and duration relative to "say" in (1). Following this, the pitch on the preceding word "I'll" was re-synthesized to match the pitch values of this word in (2), i.e. a low-dipping pitch realizing the low target of the L+H* accent. Pitch on "I'll" in the NPA condition was *also* resynthesized, overlaid with highly comparable values from another production of (1), ensuring that both "I'll"s underwent an equal amount of resynthesis, in case any artifacts from resynthesis remained that might influence perceived naturalness. Importantly, the post-target material "now" was identical across conditions, being as it was produced in (1), which was highly similar to its production in (2). In both cases it was realized as unaccented and phrase-final with a low (L-L%) boundary tone. These manipulations thus created differences in the pre-target pitch contour, as well as the duration, overall amplitude and envelope of the pre-target vowel /eɪ/, as shown in Figure 2.1. The portion that underwent resynthesis (excluding the target) was only the word "I'll".

The starting point for the creation of the target itself was a production of "ebb", produced with an H* pitch accent, as in (1) above. Because the goal was to create a target that would be appropriate for both frames and be identical across conditions, pitch and intensity for the target sound were manipulated to be the average of the nuclear accented target, as in (1), and the post-focus target, as in (2). This was intended to render the target ambiguous in terms of prominence, such that it would be interpretable as relatively prominent in the NPA context, as in (1), but also interpretable as lacking prominence when post-focus, as in (2).

F1 and F2 were manipulated by LPC decomposition and resynthesis using the Burg method in Praat (Reinisch & Sjerps, 2013; Sjerps, Mitterer, & McQueen, 2011; Winn, 2016). The formant values for each endpoint were based on model sound productions of "ebb" and "ab". The resynthesis process estimated source and filter for the starting model sound from the "ebb" model. The filter model's F1 and F2 were then adjusted to match those

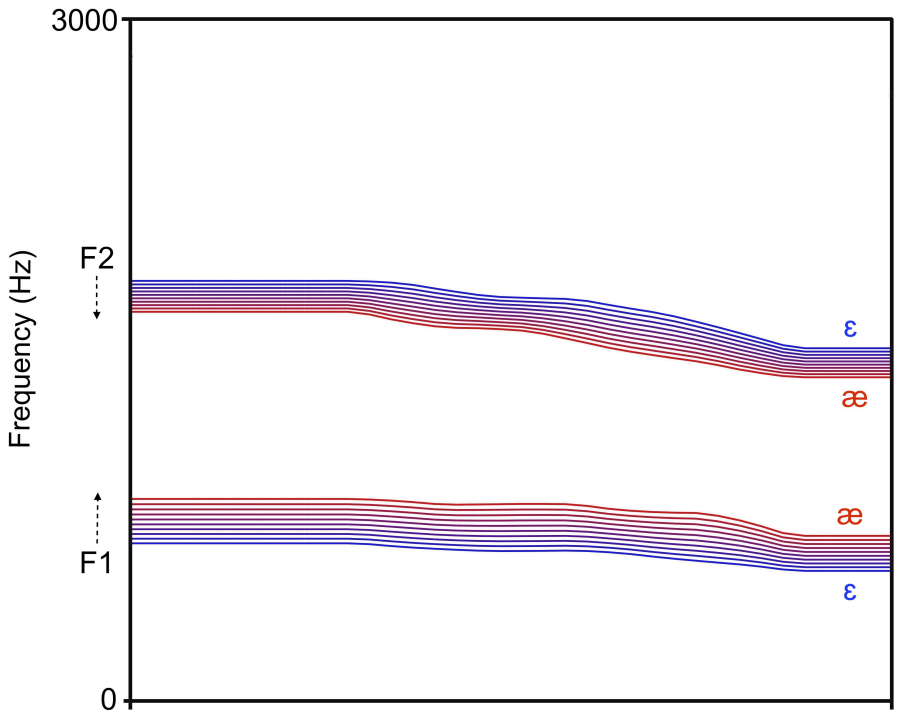

Figure 2.2: Formant tracks of the Experiment 1 continuum, with F1 and F2 shown. The outermost blue lines represent the formant values for the /ɛ/ endpoint of the continuum (mean F1 = 680 Hz, mean F2 = 1724 Hz). The innermost red lines represent the formant values for the /æ/ endpoint of the continuum (mean F1 = 838 Hz, mean F2 = 1596 Hz). The x axis is time, and is approximately 131 ms in duration, the duration of the target vowel.

of a model "ab" production. From these two filter models, 8 intermediate filter steps were created, by interpolating between these model endpoint values in Bark space (Traunmüller, 1990). Phase-locked higher frequencies from the starting base /ɛ/ model that were lost in the process of LPC resynthesis were restored to all continuum steps, improving the naturalness of the continuum. The result was a 10 step continuum ranging from model /ɛ/ to /æ/ values in F1 and F2. Intensity and pitch were invariant across the continuum. A visual representation of the F1 and F2 manipulation is given in Figure 2.2. Each continuum step was then cross-spliced into both NPA and post-focus frames, creating 20 unique stimuli in total (10 continuum steps $\times$ 2 frames).

Importantly, though the stimuli were created to cue a contrast in accentuation (i.e., phrasal, or phonological, prominence), phonetic prominence varies across conditions as well. Given that prominence perception is relative (e.g., Terken & Hermes, 2000), a target that is acoustically identical across conditions will still vary in *relative phonetic prominence* as a function of context. The target is relatively less phonetically prominent in the post-focus condition (again, as a function of context), and relatively more prominent in the NPA condition (see Figure 2.1). The stimuli thus represent variation in both phonetic and phonological prominence, a point that will be returned to in Section 2.4 as it pertains to the timecourse results from Experiment 2.

2.2.2 Predictions

As outlined above, the central prediction forwarded here is that listeners will relate formant information in the vowel to prosodic context, effectively accounting for prominence strengthening effects on formant structure. What outcome would this predict in the present experiment? Prominence strengthening in /ɛ/, following sonority expansion, would show increased F1 and decreased F2 (corresponding to jaw lowering and backing of lingual articulations). If listeners attribute these formant changes to prominence (i.e. being driven prosodically instead of signaling a phonemic contrast), they should map "strengthened" formant values (raised F1, lowered F2) to /ɛ/ more often, showing *increased* "ebb" responses in the NPA condition. In other words, in prominent contexts, listeners would interpret raised F1 and lowered F2 as being driven by prominence, not as a cue to the vowel contrast. This outcome is of course relative to the post-focus condition, in which non-strengthened variants of each vowel would be appropriate.

Given the structure of the stimuli, a competing prediction can also be made. This prediction is based on the observation that the contrast between /ɛ/ and /æ/ in American English is in part durational, where /æ/ is longer (Umeda, 1975). Because duration is a potential cue to the contrast, contextual durations in the carrier phrase may influence listeners' perception of the target sound. As shown in Figure 2.1, a longer vowel /eɪ/ precedes

the target in the post-focus condition, as compared to the NPA condition. Following standard durational contrast effects discussed in Section 1.4.3 (Diehl & Walsh, 1989; Wade & Holt, 2005), we could predict that the target should sound relatively short to listeners following longer /eɪ/ in the post-focus condition. A shorter perceived target in this condition would effectively lead to *increased* "ebb" responses in the post-focus condition, if duration is used as a cue to the contrast. Given that this effect is the opposite of the prominence effect laid out above, this can be seen as a fairly conservative test for prosodic effects, testing a case where general auditory factors (i.e., durational context effects) predict a different outcome.

2.2.3 Participants and procedure

30 participants were recruited for Experiment 1. All were self-reported native American English speakers with normal hearing, and were recruited from the UCLA student population. Each participant completed a language background questionnaire and provided informed consent to participate. Participants received course credit for their participation. The online platform that was used to control stimulus presentation in Experiment 1, and all subsequent behavioral categorization experiments (that is, experiments that did not involve eyetracking) was Appsobabble (Tehrani, 2020).

The procedure was a simple two-alternative forced choice (2AFC) task in which participants heard a stimulus and categorized it as one of two words, "ebb" or "ab". Participants completed testing seated in front of a desktop computer monitor, in a sound-attenuated room in the UCLA Phonetics Lab. Stimuli were presented binaurally via a PELTOR™ 3M™ listen-only headset. The target words were represented orthographically, each target word centered in each half of the monitor. The side of the screen on which the target words appeared was counterbalanced across participants, such that for half of the participants "ebb" was on the left, and for the other half "ebb" was on the right.

Participants were instructed that their task was to identify which word they heard by key press, where a "j" key press indicated the word on the right of the screen, and an "f" key press indicated the word on the left. Prior to the test trials, participants completed 4 training

trials. In these trials, the continuum endpoints were presented once in each prominence condition. In the subsequent test trials, each unique stimulus was presented 10 times, in random order, for a total of 200 test trials during the experiment (20 unique stimuli $\times$ 10 repetitions). Halfway through the test trials, participants were prompted to take a short self-paced break. The experiment took approximately 15-20 minutes to complete in total.

2.2.4 Results and discussion

Statistical assessment of the categorization responses in Experiment 1 was carried out using a Bayesian logistic mixed-effects regression model implemented with the *brms* package (Bürkner, 2017) in R (R Core Team, 2020). The default prior distribution, an improper uniform distribution over real numbers, was used.[6] The output of the model includes a joint posterior distribution of model parameters in addition to summary statistics for each estimated marginal distribution. In reporting the results, the estimated mean and 95% credible interval (CI) are given for each fixed effect. Evaluation of an effect's impact on categorization is carried out by considering the relevant CI, and crucially whether their interval includes zero. A 95% CI that excludes zero is taken to show that a given factor has a meaningful (i.e., credible) impact on listeners' responses.

The model was structured to predict listeners' categorization response (with "ab" mapped to 0 and "ebb" mapped to 1) as a function of continuum step and prominence manipulation, as well as the interaction of these two fixed effects. Continuum step was treated as a continuous variable, scaled and centered at 0. Prominence condition was contrast-coded, with NPA mapped to 0.5 and post-focus mapped to -0.5. The random effect structure specified in the model consisted of by-participant random intercepts and random slopes for both fixed effects and their interaction. The fixed effect estimates from the model are shown in Table 2.1. Categorization responses are plotted in Figure 2.3.

As shown in Table 2.1, continuum step impacted categorization such that as contin-

[6]The reader is referred to Bürkner (2017), Vasishth, Nicenboim, Beckman, Li, and Kong (2018) and Chodroff and Wilson (2019) for detailed descriptions of Bayesian modeling and recent application to similar data.

Table 2.1: Model output for Experiment 1, with estimates for each fixed effect, estimate error, and 95% CI. A checkmark in the rightmost column indicates that an effect is credible, i.e. that the 95% CI excludes zero.

	Estimate	Est. Error	L-95% CI	U-95%CI	credible?
intercept	0.05	0.17	-0.29	0.39	
prominence	0.83	0.28	0.27	1.39	✓
continuum	-2.57	0.28	-3.15	-2.03	✓
prominence:continuum	-0.24	0.13	-0.50	0.01	

uum step increased (i.e., became less /ɛ/-like), "ebb" responses decreased (β=-2.57, 95%CI =[-3.15,-2.03]). This is clearly visible in Figure 2.3, and expected for any such continuum. Prominence, the predictor of interest, also showed a credible effect, such that "ebb" responses increased in the prominent NPA condition (β=0.83, 95%CI =[0.27,1.39]).This is also visible in Figure 2.3, where the categorization function is shifted across prominence conditions.

This finding supports the predictions laid out above: contextual prominence, as cued by phrasal organization and intonation, shifted listeners' perception of the target vowel such that they more readily categorized a prominent target as "ebb". This finding provides new evidence for the involvement of prosodic factors in speech perception, and shows prosodic prominence plays a role in listeners' perception of formant cues. The results of Experiment 1 are also not explainable on the basis of durational contrast (cf. Mitterer et al., 2016) as discussed in Section 2.2.2 above.

Experiment 1 therefore affirmatively answers the question of whether prosodic prominence mediates segmental processing, though *how* this information is being integrated with formant cues by listeners remains an open question. Experiment 2 addresses this question, exploring the processing questions outlined in Section 1.4.4.

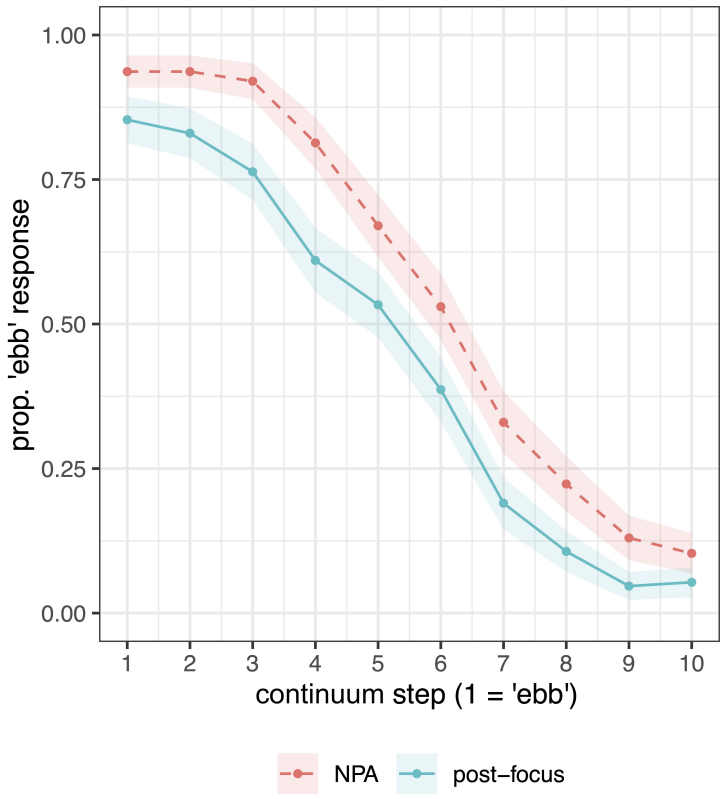

Figure 2.3: Categorization responses in Experiment 1, with the proportion of "ebb" responses plotted on the y axis, split by prominence condition and continuum step, where step 1 is the /ɛ/ endpoint of the continuum. Shading around each line shows 95% CI.

2.3 Experiment 2

Experiment 1 showed that listeners incorporated phrasal prominence into their perception of a vowel contrast, in line with how vowels are strengthened phonetically when prominent. The goal of Experiment 2 was to test when in processing this effect occurs. Two possibilities, discussed in Section 1.4.5, are considered: (1) later-stage modulation of lexical competition via prosodic analysis (Kim, Mitterer, & Cho, 2018; Mitterer et al., 2019), and (2) immediate compensation, or re-coding of a cue pre-lexically (McMurray & Jongman, 2011; Toscano & McMurray, 2015). These predictions can be further specified given the structure of the stimuli, which are the same stimuli as used in Experiment 1.

First, as a preliminary, it is important to make explicit an assumption about listeners' processing in a task in which they are categorizing a phonetic continuum. Following e.g., Newman et al. (1997), it is assumed here that in a 2AFC task in which listeners categorize a continuum, an ambiguous token on that continuum will cause listeners to activate both continuum endpoints as lexical hypotheses. Factors which contribute to an eventual decision (categorization response) can in this sense be seen as modulating the process of lexical competition between the two endpoints under consideration. The timing of modulation can be assessed by looking at how listeners' looks to a target word change over time, allowing us to test at what point various sources of information become relevant in processing (see also Mitterer & Reinisch, 2013; Toscano & McMurray, 2015).

With this in mind, we can consider two pieces of information that the stimuli used in Experiment 1 provide to listeners. Firstly formant cues, being a primary dimension for the contrast between /ɛ/ and /æ/, should, trivially, be useful to listeners in identifying the vowel. Formants in a vowel can further be characterized as an *intrinsic* cue: they are produced as part of the vowel articulation and provide temporally co-occurrent information about the vowel as it unfolds (as opposed to preceding or following it in time). In terms of the model sketched in Section 1.4.1, they should contribute to the early stages of processing (pre-lexically) for both target words, with ambiguous values activating both lexical hypotheses. Reinisch and Sjerps (2013) also showed that listeners rapidly use vowel-intrinsic spectral cues in perception, in line with this view. On this basis, we should expect the use of formants (that is, changing F1/F2 along the continuum) to rapidly influence listeners' looks to a target word. We can consider this as a sort of temporal benchmark for what counts as "rapid" in this experiment.

Experiment 1 also showed that phrasal prominence shaped listeners' perception of the /ɛ/-/æ/ contrast. As described in Section 2.2.1, the target word was acoustically identical across prominence conditions such that the prominence-lending nature of the carrier phrase was purely contextual. We can therefore describe prominence as a contextual cue to the contrast (as established in Experiment 1), which crucially precedes the target sound in time. Recall that material following the target is identical, such that all differences in the two

prominence conditions precede the target sound (see Figure 2.1).

We can now re-frame the two accounts outlined above in terms of the point in time at which both phrasal prominence and vowel formants impact processing. The prosodic analysis account and recent findings in its support (Kim, Mitterer, & Cho, 2018; Mitterer et al., 2019) make a clear prediction. Recall that in the prosodic analysis model, segmental categories (here, cued by formants) activate lexical hypotheses, while prosodic information is used to modulate lexical competition. Phrasal prosody thus should exert a later-stage influence, being integrated post-lexically. This, in relation to vowel-intrinsic formant cues, should occur at a later point in time. Listeners' use of prominence information should thus be *asynchronous* with their use of formant cues.

On the other hand, if (phonetic) prominence immediately modulates perception of the target sound via expectations generated by preceding material in the carrier phrase, and compensatory perceptual re-coding, we should expect to see an early (i.e. pre-lexical) influence of prominence context. Following Toscano and McMurray (2015), if prominence-lending context modulates the perception of formants directly, its influence should therefore be seen at the same time as the vowel-intrinsic cue. That is, prominence and formant cues should *simultaneously* impact segmental processing in its early stages, and show a similar overall timecourse.

Consider another difference implied by these predictions. In the prosodic analysis account, early stages of lexical activation should be the same across conditions, that is, listeners' use of formant cues early in processing should show veridical perception of formants that *does not vary* across prominence conditions, because phrasal prominence is not modulating perception of the formants themselves. It follows that eye-movement differences across conditions should only be apparent relatively late in processing. On the other hand, the phonetic context account predicts that early processing of formant information *should vary* across conditions, as the perception of formant values is being shaped directly by prosodic context. In this sense, looking at the early use of formant cues themselves, across conditions, may further help decide between these accounts.

These timecourse predictions are summarized in Table 2.2.

Table 2.2: Timecourse predictions for Experiment 2.

mechanism	order of cue usage	early formant processing
prosodic analysis	formants before context	the same across conditions
phonetic context	simultaneous	different across conditions

2.3.1 Materials

The materials in Experiment 2 were a subset of those used in Experiment 1. With the goal of sampling from more ambiguous stimuli (Kingston, Levy, Rysling, & Staub, 2016; Mitterer & Reinisch, 2013; Reinisch & Sjerps, 2013), the middle region of the continuum was chosen for this purpose. The method by which the Experiment 2 stimuli were selected was the same as that used in Mitterer and Reinisch (2013). First, the overall interpolated categorization function for Experiment 1 was inspected. The point at which the interpolated function crossed 50% (i.e. the most ambiguous region in the continuum) was identified. The three steps on each side of this crossover point were used in Experiment 2. This led to the selection of steps 3-8 from Experiment 1. Note that these steps are re-numbered as steps 1-6 in what follows, where step 1 in Experiment 2 refers to step 3 in Experiment 1, and so on. There were accordingly 12 unique stimuli used in Experiment 2 (6 continuum steps $\times$ 2 prominence conditions).

2.3.2 Participants and procedure

36 participants were recruited for Experiment 2 from the same population as Experiment 1. All participants additionally had normal or corrected-to-normal vision.

The paradigm used in Experiment 2 was an adaption of that used by Reinisch and Sjerps (2013), and Kingston et al. (2016). It was a visual world eyetracking task in which partici-

pants viewed an orthographic display of the target words "ebb" and "ab". The participants' task was simply to click on the word they heard. Participants' eye movements were monitored while they listened to stimuli and provided their responses. Testing was carried out in a sound-attenuated room in the UCLA Phonetics Lab.

Participants were seated in front of an arm-mounted SR Eyelink 1000 (SR Research, Mississauga, Canada) set to track the left eye at a sampling rate of 500 Hz, and set to record remotely (i.e., without a head mount) at a distance of approximately 550 mm. At the start of the experiment, participants' gaze was calibrated with a 5-point calibration procedure.

Stimuli were presented binaurally via a PELTOR^{TM} 3M^{TM} listen-only headset. The visual display was presented on a 1920×1080 ASUS HDMI monitor. During each trial, participants were first presented with a black fixation cross (60px by 60px) in the center of monitor. The target words themselves were displayed in 60pt black Arial font, with one word centered in the left half of the monitor, and the other in the right half of the monitor. The side of the screen on which the words appeared was counterbalanced across participants, though for a given participant the same word always appeared on the same side of the screen (Kingston et al., 2016; Reinisch & Sjerps, 2013). Two interest areas (300px by 150px) were defined around the target words. These were slightly larger than the printed words, to ensure that looks in the vicinity of the target words were also recorded, following e.g., Chong and Garellek (2018); Kingston et al. (2016).

The onset of the audio stimulus was look-contingent, such that stimuli did not begin to play until a look to the fixation cross had been registered. This was done to ensure that participants were not already looking at a target word at the onset of the stimulus. As soon as a look to the fixation cross was registered, the audio stimulus began, and the target words appeared simultaneously with the onset of the audio. The trial ended after participants provided a click response. The next trial began automatically after a click response was registered. At the start of each new trial, the cursor position was re-centered on the computer screen, following Kingston et al. (2016). Trials were separated by an interval of 1 second. Eye movements were recorded from the first appearance of the fixation cross until the participants provided a click response and the next trial began.

Table 2.3: Model output for Experiment 2 click responses.

	Estimate	Est. Error	L-95% CI	U-95%CI	credible?
intercept	0.09	0.16	-0.21	0.40	
prominence	0.91	0.35	0.22	1.59	✓
continuum	-1.56	0.20	-1.96	-1.17	✓
prominence:continuum	-0.13	0.12	-0.37	0.11	

There were a total of four practice trials, as in Experiment 1, with each continuum endpoint being presented in each prominence condition once. Following this, there were a total of 96 test trials; each of 12 unique stimuli was presented a total of 8 times, with stimulus presentation completely randomized. The experiment took approximately 20 minutes to complete in total.

2.3.3 Results and discussion

2.3.3.1 Click responses

Listeners' click responses (their categorization of the target word) were analyzed using a Bayesian logistic mixed-effects regression model with the same model structure as that in Experiment 1. The goal of this analysis was to confirm that listeners' categorization was influenced by prominence in this experiment and in that sense replicate Experiment 1. The model output is shown in Table 2.3, and categorization responses are plotted in Figure 2.4. As expected, increasing continuum step (becoming less /ɛ/-like) decreased clicks on "ebb" (β=-1.56, 95%CI =[-1.96,-1.17]). The prominence effect was replicated as well, whereby the NPA condition showed increased clicks on "ebb"(β=0.91, 95%CI =[0.22,1.59]). This outcome roughly mirrors the effects seen in Experiment 1, though we can note the stimuli are overall more ambiguous to listeners, as would be expected given that the central region of the continuum from Experiment 1 was used.

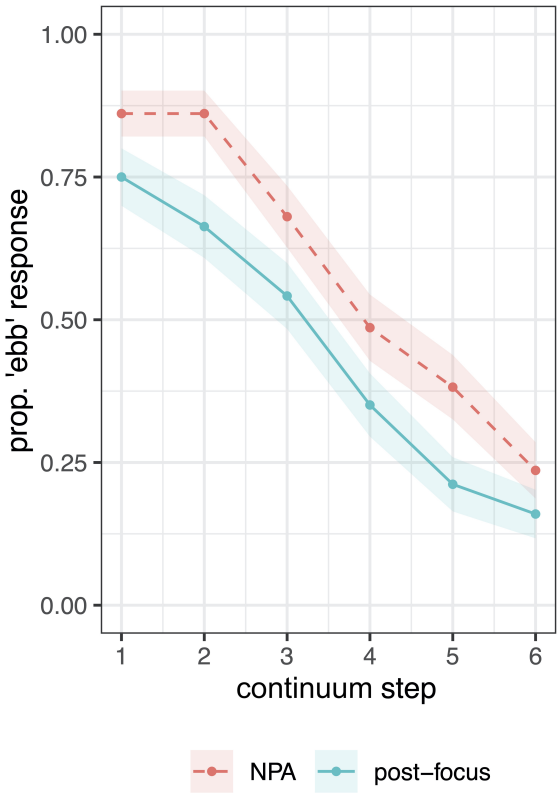

Figure 2.4: Categorization (click) responses in Experiment 2, showing the proportion of "ebb" responses on the y axis, split by prominence condition and continuum step. Note that steps 1-6 in Experiment 2 correspond to steps 3-8 in Experiment 1, as described in the text.

2.3.3.2 Eye movement data

Results for eye movement data are shown in Figure 2.5, where listeners' preference for "ebb" is plotted over time, split by continuum step in panel A, and by prominence condition in panel B. The average duration of a trial in the experiment was 1384 ms. Following Nixon, van Rij, Mok, Baayen, and Chen (2016), the analysis window accordingly spanned from 200 ms prior to target onset until 1300 ms following the onset of the target, given that effects of lexical competition have been seen to persist until this time (Dahan, Magnuson, Tanenhaus, & Hogan, 2001).

The preference measure which is represented visually in Figure 2.5 is simply the propor-

tion of looks to the "ab" interest area subtracted from the proportion of looks to the "ebb" interest area for a given point in time (with time binned by 20 ms intervals). Visually representing listeners' looks in this way allows for a normalized measure of their preference for one target over another (note the opposite preference measure would show the same information, only with the directionality inverted). Showing only the proportion of looks to "ebb", or to "ab" gives qualitatively similar results. However, it is not the case that looks to "ebb" for a given time and condition will necessarily be inversely proportional to looks to "ab" at that time (given that participants could be looking to neither "ebb" nor "ab"). The preference measure is therefore advantageous in that the effect does not vary based on whether looks to "ebb" or "ab" are being visualized.[7] With this measure, a negative preference for "ebb" accordingly corresponds to a preference for "ab".

As can be seen in both panels of Figure 2.5, this preference measure is zero at the beginning of the analysis window, indicating that listeners do not have an immediate preference for either target prior to the onset of the target word, or at the onset of the target word. Given that it takes approximately 200 ms to program a saccade (Fischer, 1992; Matin, Shao, & Boff, 1993), this timing delay should be kept in mind in the discussion of timecourse results that follows. In the top panel of Figure 2.5, we can see that over the course of a trial a preference for "ebb" develops on the basis of continuum step, such that listeners show the strongest preference for step 1, the most "ebb"-like on the continuum. Listeners show a graded preference based on continuum step, such that at step 6, the most "ab"-like, they show the strongest preference for "ab", with other steps showing intermediate degrees of preference. This suggests broadly that listeners used formant cues online to determine the identity of the target word, which is not surprising. We can see an analogous split in looks based on the prominence manipulation, shown in panel B of Figure 2.5. A prominent target, in the NPA condition, lead listeners to develop a preference for "ebb" over the course of trial. This effect is clearly smaller than that of continuum step, but nevertheless suggests a robust role for the prominence manipulation, in that there is, overall, a reliable separation in looks

[7] An exploratory analysis found that using a non-transformed preference measure, modeling looks only to "ebb" or "ab", resulted in essentially the same results, as would be expected (cf. Kingston et al., 2016).

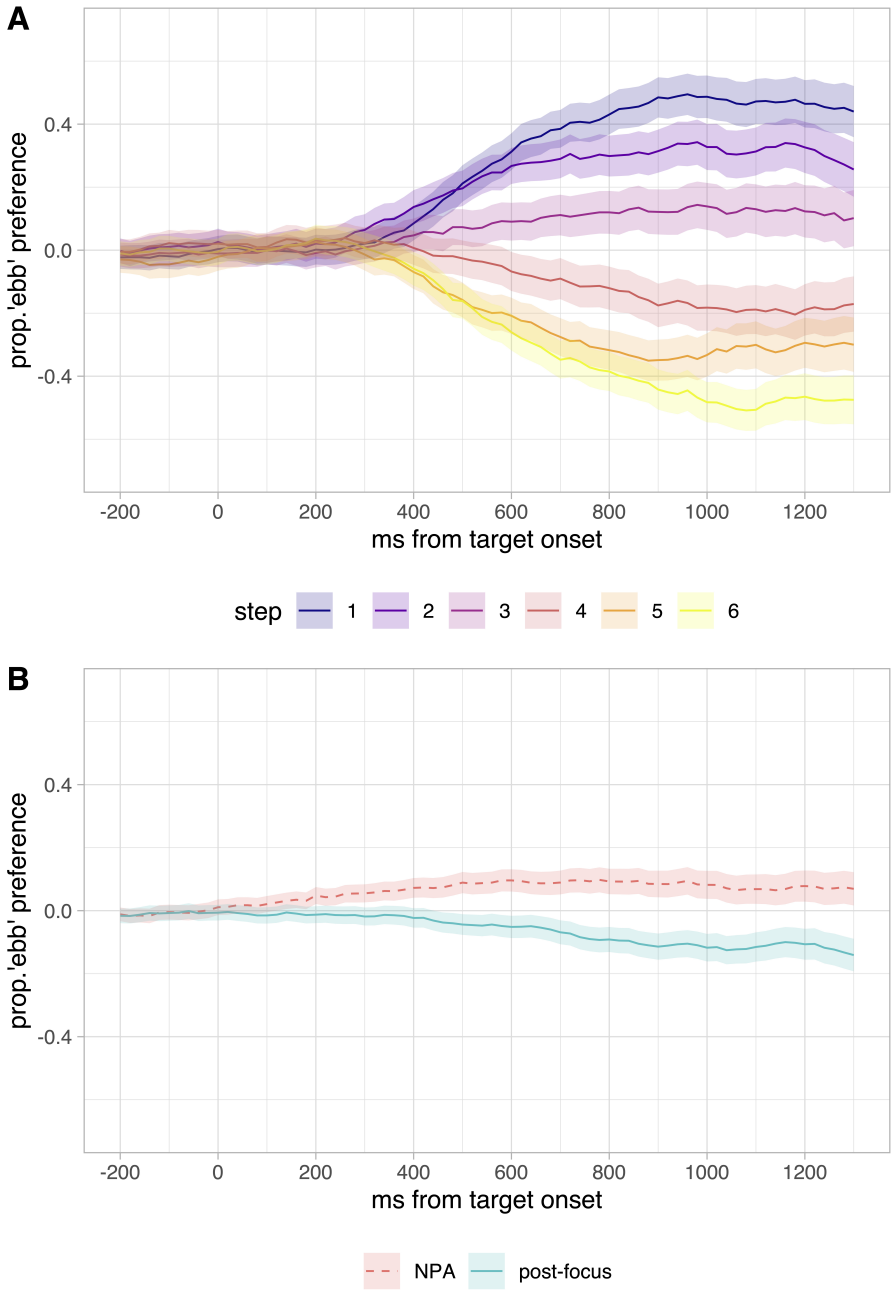

Figure 2.5: Eye movement data for the effect of continuum step (panel A), where step 1 is the /ɛ/ endpoint of the continuum, and prominence manipulation (panel B), in Experiment 2. The x axis shows time ranging from -200 to 1300 milliseconds from the onset of the target word. The y axis shows the proportion of looks to "ebb" minus the proportion of looks to "ab" (see text). Confidence regions around each line represent 95% confidence intervals, calculated from the raw data.

in the analysis window. Notably, divergence on the basis of prominence appears to start early: looks begin to pull apart fairly robustly, around 250 ms from the onset of the target. However, the effect appears to grow slowly, and does not reach a stable maximum until later in time.

The timecourse of both of these effects was assessed by a General Additive Mixed Model (GAMM). GAMMs have been applied in various analyses of visual world eyetracking data and present a powerful tool for modeling dynamic and nonlinear effects over time, especially for data with high degrees of autocorrelation, like eye movement data. GAMMs model dependencies via smooth functions: linear and parabolic functions of varying complexity, which include a pre-specified number of base functions. Fixed parametric terms in the model can also be used to model effects in an overall analysis window as in linear mixed-effects regression models.[8]

The dependent measure in the analyses reported here is a log-transformed normalized preference measure, using the same method as that used by Reinisch and Sjerps (2013), who employed a similar 2AFC visual world paradigm. This measure was calculated as log-transformed looks to "ebb" minus log-transformed looks to "ab", using the empirical logit (Elog) transformation given in Barr (2008):

$$Emprical\ logit = ln\left(\frac{y + 0.5}{n - y + 0.5}\right)$$

In this transformation, n is the total number of samples in a given time bin and y is the number of samples for a given interest area ("ebb" or "ab"). As mentioned above, an alternative would be to model looks to just one of the target words, as in e.g., Kingston et al. (2016). A preliminary analysis with both of these non-normalized measures modeling looks to "ebb" or "ab" showed qualitatively similar results to the preference measure, though they are not reported on further here.

[8]The reader is referred to Nixon et al. (2016) and Zahner et al. (2019) for discussion of advantages of GAMMs in modeling visual world eyetracking data, and to Sóskuthy (2017) and Wood (2017) for a more general overview of GAMMs.

The model was fit in R using *itsadug* and *mgcv* (van Rij, Wieling, Baayen, & van Rijn, 2016; Wood, 2017). Parametric terms in the model predicted the preference measure as a function of (scaled and centered) continuum step and prominence condition, which was contrast-coded as in previous experiments. The smooth terms in the model included a non-linear interaction term of continuum by prominence condition, over time, allowing us to assess how listeners' preference for a target develops over time as a function of both these factors. This was modeled with the *te()* function in *mgcv*, which includes main effects and interaction terms. Random effects were modeled using factor smooths, which are analogous to random slopes and intercepts in other mixed models. Factor smooths were fit to by-participant trajectories in each prominence condition, allowing for the possibility that participants were impacted differently by the prominence manipulation.[9] Both added smooth terms significantly improved the model fit, as assessed by comparing models with the *CompareML()* function. Importantly, the inclusion of condition in the *te()* term improved the model fit significantly ($\chi^2(5)$=180.27, p<0.001).[10] This suggests that listeners' use of formant cues over time varies across conditions (i.e. in addition to overall variation in height of the trajectories captured by the parametric term). This point will be returned to later. The default number of basis functions (knots) was employed for each smooth term, and this was observed to provide a good fit to the data by inspecting the k′ scores and k-indices in the model using the *gam.check()* function in *itsadug*.

Following Nixon et al., 2016 and Zahner et al., 2019, the timecourse data was downsampled to 50 Hz (20 ms bins), allowing for a fairly granular timecourse assessment, while reducing autocorrelation among successive bins. Because some residual autocorrelation remained, following Nixon et al. (2016) and Zahner et al. (2019), an AR1 error model was employed after inspection of the baseline model, as it greatly reduced autocorrelation as

[9]These factor smooths were shown to provide a better model fit than trajectories that were only by-participant, as assessed by comparing fREML scores using the *CompareML()* function in *itsadug*, including more complex factor smooths both increased fREML and decreased AIC.

[10]This comparison was carried out by comparing model scores using the *CompareML()* in *itsadug*, as in Nixon et al. (2016). The original model was compared to one in which prominence condition was removed from the three way interaction.

compared to the non-AR1 variant (Nixon et al., 2016; Sóskuthy, 2017).[11] Note that the numerical output of a GAMM is not particularly useful for evaluating an effect; results are often best assessed by visualizing aspects of the model fit; as stated by Zahner et al. (2019, p 85): "GAMM model outputs alone are not sufficient for the interpretation of the results, effects only become obvious through visualization" (see also Nixon et al., 2016; Sóskuthy, 2017; Wood, 2017). The full model output is contained in Table B.1, in Appendix B.

The parametric terms in the model, which represent the overall effect in the full analysis window, indicate that both prominence condition and continuum step had an effect on listeners' preference for each target word. In line with what can be seen in Figure 2.5, increases in continuum step (becoming less "ebb"-like) decreased listeners' "ebb" preference (β= -1.63, t=-18.04). At the same time, the prominent NPA condition showed a marginal influence in the analysis window as whole, increasing listeners' "ebb" preference (β= 0.44, t=1.91). The parametric terms thus confirm the manipulations are influencing looks within the analysis window as expected given our observation of the raw data, but they do not tell us about the timecourse of each effect.

To assess the timing of the effect of phrasal prominence and the effect of changing F1 and F2 along the continuum, differences between smooths of interest (over time) were inspected (Sóskuthy, 2017; Zahner et al., 2019). By observing when the difference between two relevant smooths (comparing across conditions of interest) becomes reliable, we can assess when these trajectories diverge, and thus when listeners' eye movements are first impacted by the conditions which are being inspected, and more generally, how the effect changes over time.

The effect of continuum step was assessed by visualizing the difference between smooths for the two continuum steps which spanned the most ambiguous region in the continuum (steps 3 and 4).[12] Given that each step has its own trajectory, pairwise differences between

[11] AR1 models assume that neighboring observations in e.g., a time series, are correlated such that the error in one time bin (in this case) is in part dependent on the error in adjacent bins. Assuming correlated errors in parameter estimation helps remove correlations among residuals; see e.g., Baayen, van Rij, de Cat, and Wood (2018) for more information.

[12] The divergence estimate shown in panel A of Figure 2.6 is for the trajectories for these steps in the post-focus condition (as collapsing across conditions is not possible); the effect in the NPA condition showed a comparable timecourse.

steps may not necessarily be the same. As can be observed in Figure 2.5, this area between steps 3 and 4 shows the most robust pairwise difference, and the processing of acoustic information at these steps makes a relevant comparison to the effect of prominence, which was calculated for the middle of the continuum, described below. This thus represents how quickly formant information is used to distinguish more ambiguous vowels, though notably the estimates for the rest of the pairwise differences between steps only differed by 10-20 ms.[13] This effect is shown in panel A of Figure 2.6.

To assess the point in time at which phrasal prominence shows a robust effect on the preference measure, the difference between smooths for each prominence condition was visualized over time. The continuum step at which this divergence was calculated was set to be the scaled value of 0, between step 3 and 4 on the continuum. This represents the most ambiguous region on the continuum, where context should exert the strongest effect, and therefore where we should expect to see the earliest effect of prominence, a conservative test given the prediction that the effect will be later in processing. This effect is shown in panel B of Figure 2.6.

As can be seen in Figure 2.6, the points in time at which formants and phrasal prominence impact listeners' preference reliably are asynchronous, with the effect of the continuum (formants) preceding the effect of prominence (see figure caption). The model estimates that looks diverge based on the continuum at 270 ms following the onset of the target vowel, a clearly early effect considering the 200 ms required to program a saccade. The model further estimates that looks diverge based on phrasal prominence at 482 ms following of the target. This is shown for both effects in Figure 2.6 when CI for the model estimate do not include zero, indexed by a dashed vertical line.[14] Another possible way of operationalizing the effect

[13]Following Maslowski et al. (2020), an alternative operationalization of the effect would be to compare the two steps which are most different acoustically (i.e. steps 1 and 6). This comparison yielded a similar though slightly earlier effect, with divergence estimated at 258 ms after target onset.

[14]This timing asynchrony was also seen in a more traditional moving window analysis, not included here. In that analysis, time was binned into 100 ms windows and a linear mixed-effects regression on log-transformed preference measures was run in each. Continuum step began to have a significant effect in the 300-400ms window. Phrasal prominence began to have a significant effect in the 500-600 ms window, though notably the prominence effect approached significance earlier in time, in similar fashion to Kim, Mitterer, and Cho (2018).

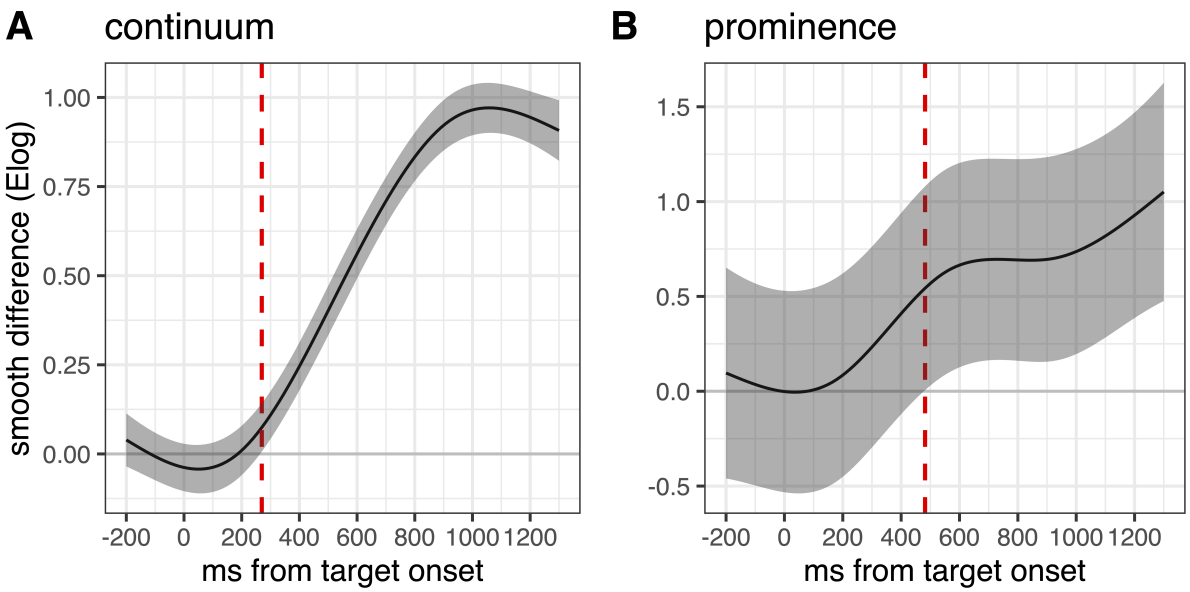

Figure 2.6: Difference smooths (i.e. differences between smooths of interest, as described in the text) for the effect of continuum step (panel A), and the prominence manipulation (panel B), in Experiment 2. The x axis shows time in the analysis window, the y axis shows the difference between smooths in listeners' log-transformed preference measure (see text). Smooths are surrounded by 95% CI, and the red dashed vertical lines index when in time CIs exclude zero, that is, when the difference between smooths becomes reliable (i.e. reliably non-zero). Note the y axes are different in each panel.

of prominence would be to calculate the prominence effect at each step on the continuum and take the average. In this case, we would not be inspecting the most ambiguous region of the continuum, where processing would be expected to be early, but instead the effect for the continuum more holistically. As expected, this estimate yielded a robustly later effect: 720 ms following the target onset. This measure further strengthens the claim that the prominence effect is overall later in processing.

Considering these divergence times alone, this outcome is consistent with the asynchrony predicted by prosodic analysis (see Table 2.2). F1 and F2 values should lead to early and immediate looks to a target, and prosody (prominence) should mediate lexical selection later in processing (Cho et al., 2007; Mitterer et al., 2019). The timing for the effect of

continuum step is consistent with previous work that shows vowel-intrinsic formant cues are used rapidly in processing (Reinisch & Sjerps, 2013). The timing of smooth divergence for phrasal prominence is clearly later, considering it follows the effect of continuum by over 200 ms, and especially considering all relevant differences in context precede the target in time, as discussed above.

Another recent study, Maslowski et al. (2020), offers a relevant comparison to the present data. Maslowski et al. explored how non-adjacent preceding speech rate before a target influenced processing of vowel duration (as cue to a vowel length contrast in Dutch). These sorts of distal rate effects are argued to operate early in auditory processing (Bosker, 2017; Bosker et al., 2017; Reinisch & Sjerps, 2013), and indeed the authors found essentially synchronous use of distal speech rate and vowel duration (lining up with the effect of rate and VOT from Toscano & McMurray, 2015, discussed in Section 1.4.5). The authors also manipulated global speech rate (that is, speech rate variation over the course of the entire experiment). Global rate effects are argued to operate later in processing, as they are sensitive to talker identity and can be overridden by other effects (Maslowski, Meyer, & Bosker, 2019; Reinisch, 2016). The authors found global rate effects showed a clear delay in processing relative to preceding (stimulus-internal) rate effects and the effect of intrinsic vowel duration, reaching significance roughly 250 ms after the effect of vowel duration itself. This relative timing difference observed by the authors is analogous to the asynchrony observed here between formant cues and prominence. The similarity between these two findings is accordingly a delay between the effect of higher-level processing (by hypothesis, prosodic analysis in the present results) and an intrinsic cue, which is used rapidly. In reference to the "order of cue usage" prediction in Table 2.2, we can therefore take these results as offering clear support for the prosodic analysis account.

Though these results would therefore suggest the effect of prominence is relatively delayed in processing, we can see that the difference between smooths begins to increase well before this point in time (visible in panel B of Figure 2.6). This is also apparent in the raw data, shown in Figure 2.5. Looks begin to diverge based on prominence condition earlier in time, and the effect grows slowly until it stabilizes later, roughly when the effect becomes

significant as assessed by the difference smooth analysis. This suggests a possible subtle and earlier influence of prominence in processing. We can test for this influence by observing if listeners' early processing of formants varies across conditions, as outlined in Section 2.3. That is, if prominence does indeed exert an earlier influence in processing, we would expect it to shape listeners' early use of formant cues. To explore this possibility, the non-linear interaction between continuum step, time, and prominence condition (which would evidence an asymmetrical influence of continuum step across conditions, over time) was inspected.[15]

To assess this interaction between continuum step, prominence condition, and time, three dimensional topographic surface plots were inspected. These plots represent the effect of continuum step (as a continuous variable on the y axis) over time (on the x axis). The dependent variable (listeners' Elog-transformed "ebb" preference) is represented on a gradient color scale. In Figure 2.7, two such plots, split by prominence condition, represent how listeners' preference changes over time and across the continuum, in each prominence condition. A value of zero (in the middle of the color scale) represents no preference, while a positive value (closer to yellow on the color scale) represents a preference for "ebb". A negative value (closer to purple on the color scale) represents a preference for "ab". Shading on the surface shows locations where listeners' preference is not significantly different than zero, with 95% CI.

One general pattern to note is that listeners do not show a preference early in the analysis window (shown by shading on all of the surface prior to approximately 300 ms). As time progresses, listeners develop graded preferences based on continuum step (as in Figure 2.5). At the end of the analysis window, there is a range of preferences: a strong "ebb" preference at step 1 on the continuum, and a strong "ab" preference at step 6. Note too that, generally speaking, the middle region of the continuum never attains a significant preference in either panel: that is, the model finds that the ambiguous region of the continuum remains ambiguous even at the end of the analysis window, shown by the shaded area persisting until the

[15]Recall that the presence of condition in the *te()* term in the model was shown to significantly improve the model fit ($\chi^2(5)=180.27$, $p<0.001$), suggesting that prominence effects are indeed interacting with listeners' perception of the continuum.

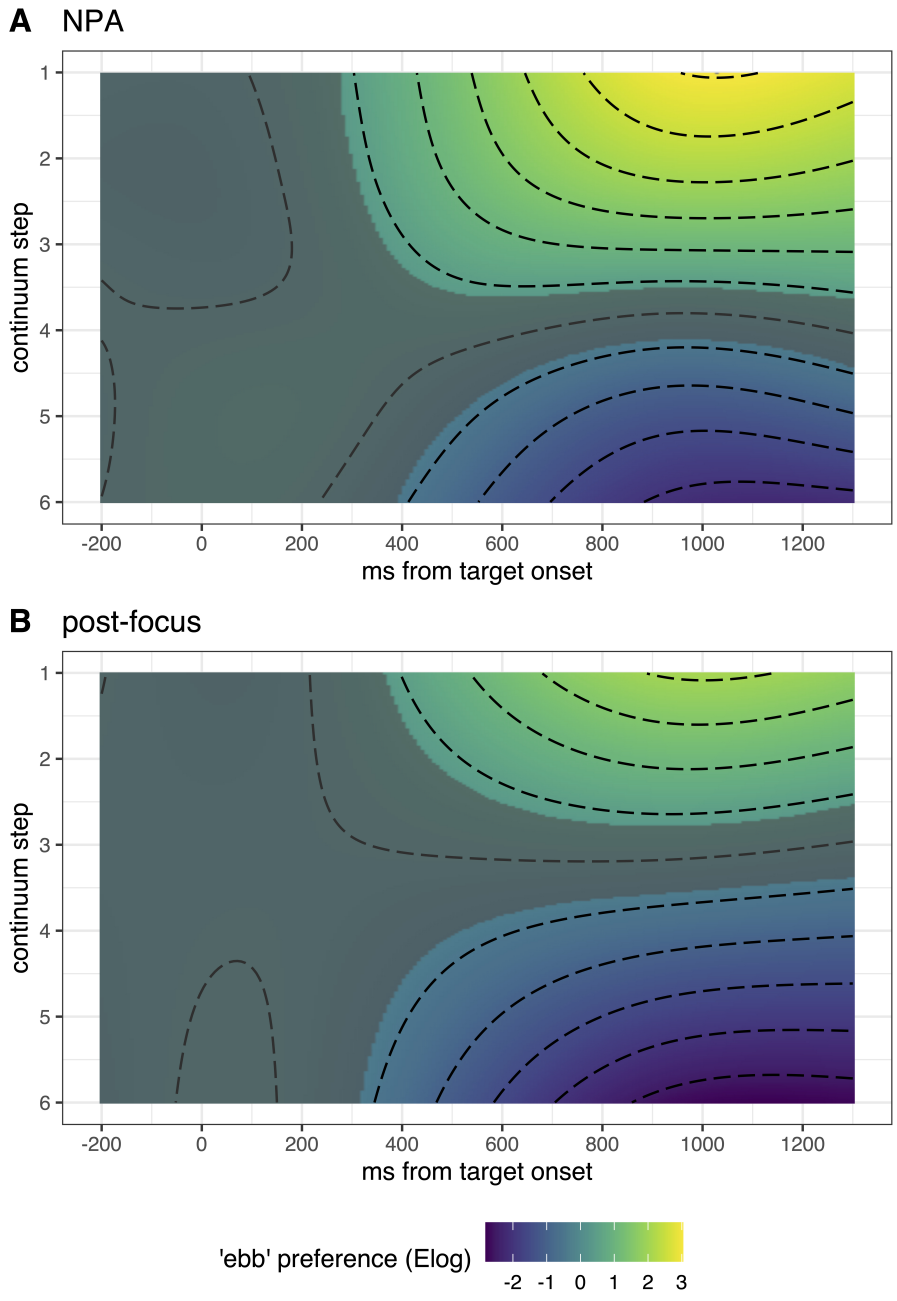

Figure 2.7: Topographic surface plots showing the effect of continuum step (y axis) over time (x axis), split by prominence condition (labeled above each panel). The color scale represents listeners' degree of "ebb" preference. Shading on the surface (the darker color that covers the leftmost portion of the surface entirely) represents locations on the surface for which the preference measure is not significantly different than zero, with 95% CI. Dotted lines show landmarks on the surface.

end of the trial in both prominence conditions.

With this in mind, we now can assess the impact of prominence on listeners' use of the continuum over time (i.e. the non-linear interaction between continuum step, time and prominence condition which contributed significantly to the model fit). The interaction is evident in observing (1) the coloration of each panel A and B, and (2) the shape and position of the shaded area showing points on the surface for which listeners' *did not* have a preference for either target. In terms of coloration, note the color scale used in both panels is shared by them, that is, the same color on each panel would reflect the same degree of "ebb" preference. We can see that each panel overall occupies different color spaces, with the NPA condition showing a stronger "ebb" preference (more yellow on the plot), and the post-focus condition showing a stronger "ab" preference (more purple on the plot). In other words, acoustically identical continuum steps are perceived as more "ebb"-like or "ab"-like a as function of prominence context. This is unsurprising, given that we see a divergence in looks based on prominence, as shown in Figure 2.5. We can note these differential preferences start to develop early in time, that is, the shape of the surfaces is different prior to 400 ms in the analysis window (this can be seen by looking at the dashed lines on the surface).

Moreover, we can note that the shaded areas (where listeners do not have a significant preference for either target) differ in how they occupy space in the surface. They also differ crucially in which steps on the continuum show a preference first, within a given panel. This is particularly clear in the NPA condition: the shaded area is asymmetrical such that more "ebb"-like steps (steps 1-3) show a significant preference (i.e. shading disappears) earlier in the analysis window, as compared to "ab"-like steps (steps 4-6). In other words, the earliest point at which listeners look to a target is influenced by phrasal prominence: the NPA condition facilitates early looks to "ebb", while it takes listeners longer to initiate looks to "ab". The opposite is true in the post-focus condition, though the pattern is less pronounced. This indicates that even in the earliest stages of processing (i.e., when listeners first show any significant preference for a target word) prominence is shaping how listeners use formant cues. As noted above, if prominence were *only* a later stage influence, we should expect the shape of the surfaces to be the same early in the analysis window. This is clearly not the

case. Also of note is the observation that in the NPA condition, the overall shaded portion of the surface is slightly smaller (approximately 48% of the surface is shaded in the NPA condition, 52% is shaded in the post-focus condition). This shows that listeners looked to a target more quickly in the NPA condition such that spaces on the surface remain ambiguous for less of the analysis window. This is tangential to the main question at hand but suggests that phrasal prominence, like lexical prominence, helps facilitate lexical processing (Cooper, Cutler, & Wales, 2002; Cutler et al., 1997; Cutler & Norris, 1988).[16]

Additionally, the surface plots show that prominence condition also influences which stimuli are perceived as ambiguous by listeners. This is apparent in looking at the vertical positioning of the shaded region, particularly the narrow portion that persists throughout the analysis window. The regions along the continuum which show no preference in looks vary based on prominence condition, starting early and persisting throughout the analysis window. This is another pierce of evidence that the prominence manipulation is shaping listeners' perception of formant cues directly. Inspection of the surface plots therefore supports a difference in early formant processing across conditions, lining up with the "phonetic context" prediction in Table 2.2.

As a way of synthesizing the two findings obtained from the divergence measures (Figure 2.6) and the surface plots (Figure 2.7), we can visualize and compare the effect of continuum step and phrasal prominence as a function of when each effect reaches its respective maximum, similar in spirit to analyses in Reinisch and Sjerps (2013) and Toscano and McMurray (2015). This was operationalized by looking at the timecourse of the difference between smooths for each effect, normalized by its minimum and maximum (i.e. the range-normalized smooth difference). These effect estimates are shown in Figure 2.8, corresponding to the smooth differences shown in Figure 2.6.

This normalized comparison allows us to inspect how each effect grows and changes over time, relative to its maximum and minimum (here for 200 ms after target onset and onward

[16]This advantage in the NPA condition exists even though the vowel preceding the target is longer in the post-focus condition (262 ms as compared to 200 ms in the NPA condition), giving listeners more time to compute the prosodic structure of the phrase as it unfolds.

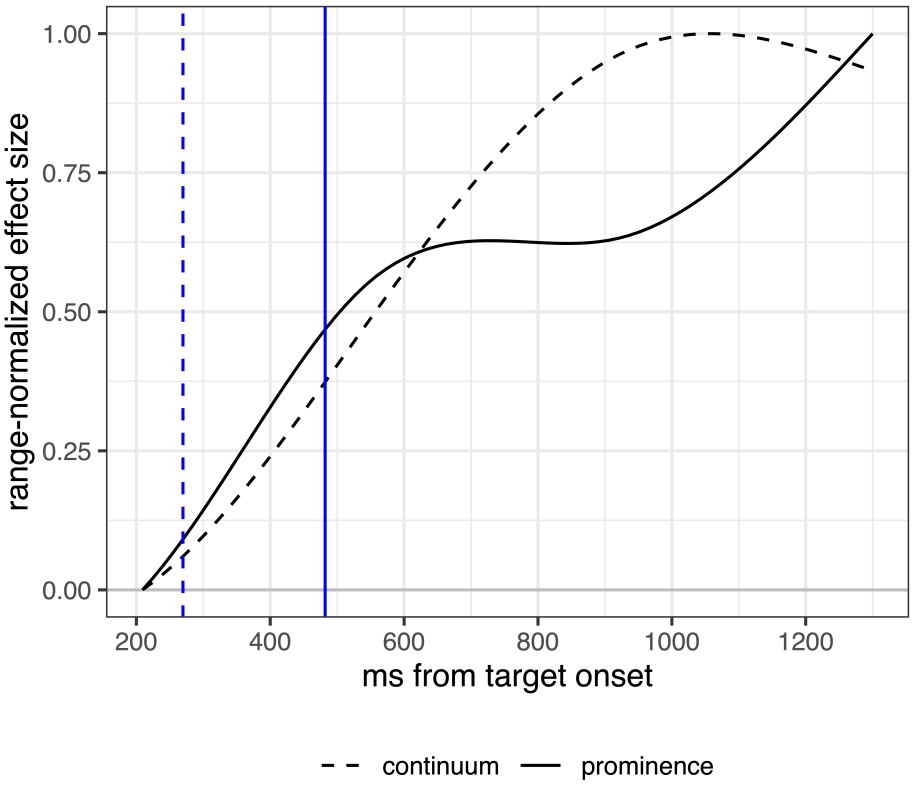

Figure 2.8: Range-normalized smooth differences, as a proportion of maximum effect, from 200 ms onward following the target onset, corresponding to the difference smooths in Figure 2.6 (see text). The solid line represents the effect of phrasal prominence, the dotted line represents the effect of the continuum. Blue vertical lines indicate when an effect becomes significant as assessed by the smooth divergence analysis, as shown in Figure 2.6.

given this delay in programming a saccade). As can be seen in Figure 2.8, the effect of continuum grows steadily, peaking around 1000 ms following the target (cf. Reinisch & Sjerps, 2013). The prominence effect is clearly different: it grows in tandem with the effect of the continuum, though growing slightly faster up until about 600 ms after target onset. While the effect of continuum continues to grow fairly linearly until its maximum, the effect of prominence actually decreases slightly, then resumes increasing around 900 ms from the onset of the target. Note that the target vowel itself is only 131 ms in duration, and listeners need only a fraction of that time to recognize the vowel based on its spectral structure as

evidenced from the timing of the continuum effect (cf. Reinisch & Sjerps, 2013). This means listeners have clearly processed the information in the vowel at e.g., 600 ms from its onset, and yet prominence information is still exerting a slow-growing and delayed effect, consistent with later stage modulation of lexical competition.[17]

We can accordingly summarize the timecourse results in Experiment 2 as the following: phrasal prominence causes looks to diverge from one another at a point that is later in time, compared to changing F1 and F2 along the continuum. At the same time, prominence actually shapes the use of formant cues earlier in time, though these effects are subtle enough not to cause overall divergence to occur early. The timecourse results thus support a multi-stage influence of prominence: one that begins early in fine-tuning formant perception, but is reinforced relatively slowly over time (in comparison to the influence of formants). This itself is somewhat visible in the raw eye-movement data in Figure 2.5: the differences between prominence conditions start early, and grow slowly over time to reach a relatively stable maximum around 600 ms from target onset. Implications of this timing outcome are discussed below, motivating the experiments carried out in Chapter 3.

2.4 General discussion

The two experiments presented in this chapter offer new insight into how listeners make use of phrasal prominence in their perception of vowel contrasts. Experiment 1 showed that listeners adjusted their perception of a vowel contrast on the basis of contextual prominence, cued by preceding pitch, duration and amplitude. In one condition, the target bore implied nuclear prominence ("implied" because the target itself did not change across conditions), and in another it was post-focus. This manipulation showed a clear effect on vowel perception: listeners modulated their categorization based on how contextually prominent the target was. In line with research that has tested how prosodic boundaries modulate perception of segmental contrasts (Kim, Mitterer, & Cho, 2018; Mitterer et al., 2019; Steffman, 2019b;

[17]This timecourse finding can be compared to a pattern observed by Kim, Mitterer, and Cho (2018): they found their AP phrasing manipulation generated small adjustments in looks early in processing, though it was subsequently "[...] weakened in the middle of the processing but reinforced later" (p 19).

Steffman & Katsuda, 2020), this finding supports a model where listeners integrate prosodic context in segmental contrast perception. In testing phrasal prominence, Experiment 1 offered an extension of past studies, and predicts that further work looking at prominence-marking prosodic features in this vein should expect to see similar compensatory effects in segmental perception.

Experiment 2 tested the timecourse of these effects in a visual world eyetracking task. The emergent pattern was complex. Formant cues were used rapidly, as would be expected. However, phrasal prominence showed, overall, a delayed influence, as measured by the point in time at which listeners' looks in each prominence condition diverged from one another. This pattern, taken by itself, is wholly consistent with the prosodic analysis model proposed by Cho et al. (2007), wherein formant cues activate lexical hypotheses, and phrasal prosody is integrated later via lexical competition (see Table 2.2). However, a subtler early influence of prominence (as shown in Figures 2.7 and 2.8), which shapes the earliest stages of formant use, suggests more nuance is needed.

Recall the discussion in Chapter 1, which made a distinction between phonological prominence related to phrasal organization, and phonetic prominence, which might derive from processing of various prominence-lending phonetic cues. As discussed in Section 2.2.1, though the target was acoustically identical across prominence conditions, its relative phonetic prominence varied. Being preceded by narrow focus marking in the post-focus condition, the target was relatively quiet, short in duration and low in pitch, as compared to the material that preceded it (see Figure 2.1). This was not the case in the NPA condition. In this sense, listeners' perception of acoustic/phonetic prominence of the target surely varied across conditions. An immediate effect of prominence in processing would accordingly reflect listeners' incorporation of phonetic prominence in their perception of formant cues. The timing of this effect is clearly early in line with general compensatory processes described in Toscano and McMurray (2015), and McMurray and Jongman (2011). Prominence effects that immediately guide formant perception are accordingly hypothesized *not* to originate from prosodic analysis (i.e., parsed out prosodic structural organization) but rather perception of acoustic/phonetic prominence, where the context-dependent perceived prominence

of the target sound shapes cue usage. This, by hypothesis, presents one measurable way in which listeners' processing of prominence information differs from prosodic boundary processing. This proposed difference stems from the notion that boundary processing is more tightly tied into the overall prosodic organization of an utterance, following Cho et al. (2007) and Mitterer et al. (2019). Prominence perception, being multidimensional in nature, is not strictly linked to the computation of (phonological) phrasal prosody.

In this light we could take the effects seen in Experiment 2 to show a multi-stage influence of phrasal prominence, broadly consistent with two-stage models of context effects in speech processing (Bosker et al., 2017; Maslowski et al., 2020; Reinisch, 2016). At the pre-lexical level of processing, acoustic/phonetic prominence shapes how formant cues are used and therefore factors into the earlier stages of lexical activation. Subsequently, lexical hypotheses are integrated with a parsed phrasal prosodic structure in prosodic analysis. This structure encodes (phonological) prominence information, and reinforces the effect, leading to more robust influences late in processing. Previous work in support of this general idea shows that many other contextual effects (e.g., speech rate, spectral context) are integrated rapidly in perception, reaching a stable maximum just following segment-internal information (Mitterer & Reinisch, 2013; Reinisch & Sjerps, 2013). The prominence effect in Experiment 2, on the other hand shows a much more delayed timecourse overall. The effect starts early, but does not stabilize until later, as shown in Figure 2.8. Given that lexical competition has been shown to persist until the later portion of the analysis window used here (Dahan et al., 2001; Salverda et al., 2007), the larger effect evidenced at these later points in time suggests prosodic context is being integrated therein, following prosodic analysis.

As a test for this claim, experiments in Chapter 3 explore another possible contextual cue to prominence: glottalization. As discussed in Section 1.3.4, in American English, glottalization in vowel-initial words has been argued to be a manifestation of prominence strengthening (Dilley et al., 1996; Garellek, 2013, 2014). Glottalization thus constitutes prominence marking that is highly localized (e.g., immediately preceding or co-occurring with a prominent vowel) in comparison to prominence conveyed by intonational tunes and accentual configuration within a phrase. Further, though it may be a manifestation of phrasal prominence,

it does not always co-occur with accentuation (Dilley et al., 1996). As such glottalization is a cue that varies *within* phonological prominence categories, and therefore can be seen as a more strictly phonetic (or at least highly localized) prominence cue. The experiments described in Chapter 3 accordingly test if glottalization is exploited by listeners as a cue to prominence, and how it is integrated with formants in processing. Comparing the on-line effects of glottalization and phrasal prominence will help confirm if the effect seen in this chapter is strengthened later in time due to the influence of phrasal prosodic structure, with glottalization serving as a comparison for a non-phrasal, but nevertheless contextual, prominence cue.

CHAPTER 3

Glottalization as a cue to prominence

3.1 The experiments in this chapter

Two experiments contained in this chapter explore how a glottal stop ([ʔ]) influences the perception of vowel contrasts, testing the hypothesis that it serves as a prominence cue. As outlined in Section 1.3.4, one apparent manifestation of prominence strengthening is glottalization at vowel onset in vowel-initial words (Dilley et al., 1996; Garellek, 2013, 2014). Observing if and how this pattern impacts perception can accordingly be seen as a test for the claim that glottalization serves a prominence-marking function. Experiment 3 tests this idea. An additional goal of this chapter is to compare the processing of [ʔ] to the phrasal prominence manipulation in Experiment 2. Testing if the timecourse of these effects differs, and if so how, will help inform our theory of prominence processing, as discussed in Chapters 1 and 2. Experiment 4 addresses these processing questions.

This chapter also includes consideration of the possible influence of an effect unrelated to prominence in vowel perception, namely *spectral contrast* (e.g., Holt et al., 2000; Stilp, 2020). Given that glottalization, and particularly [ʔ], introduces discontinuities, and temporal separation for formant trajectories in time, the impact of neighboring spectral characteristics on vowel perception is a relevant issue. This is discussed in Section 3.3.

3.2 Glottalization as a cue to prominence

As described in Chapter 1, glottalization in vowel-initial words in American English can be viewed as a form of prominence strengthening (Garellek, 2013, 2014). Some key points from

Section 1.3.4 are reviewed here.

One view of glottalization in American English is that it serves a more general function encoding prosodic boundaries and prominence (Dilley et al., 1996; Pierrehumbert & Talkin, 1992). However, a tighter link between glottalization and prominence is assumed here, in light of Garellek (2013, 2014), who showed that phrase-initial position was generally associated with *breathier voicing*, not glottalization. In light of this, glottalization is supposed to occur on phrase-initial vowels as way of counter-acting phrase-initial breathiness and strengthening spectral structure in a vowel (Garellek & Keating, 2011; Gordon & Ladefoged, 2001), which may help make cues to vowel quality more perceptible. Garellek (2014) also found that phrase-initial sonorants did not undergo this sort of strengthening, speaking against the possibility of the observed effect on vowels being the general byproduct of more forceful articulation (cf. Fougeron, 2001; Fujimura, 1990), in which case it should impact voiced segments uniformly. Moreover, phrase-level prominence, independent of prosodic boundaries, showed increased vocal fold contact, evidencing glottalization. Following the definitions given in Garellek (2013), at one level a glottal stop can be seen as an abstract articulatory target, which is realized in various "lenited" forms, such as glottalized voice quality. A glottal stop can also be an actual stop, [ʔ], made with a sustained closure of the vocal folds. This latter case can be referred to as a "full glottal stop". If we assume that these various realizations are linked to the same category, we can expect that a full glottal stop is also a possible manifestation of prominence strengthening, though Garellek (2014) looked only at voice quality measures and did not analyze the presence/absence of a full glottal stop. Data from Dilley et al. (1996) further support this: the presence of accent, including in phrase-medial position, was shown to increase the rate at which speakers produced word-initial glottal stops (and glottalization more generally).

The experiments in Chapter 2 showed that prominence at the level of the phrase (conveyed by contextual changes in pitch, duration, and amplitude) shifted listeners' perception of a target vowel, in line with how vowels are modulated acoustically by prominence (via sonority expansion). The same logic can be applied to prominence cued by glottalization. In observing if listeners shift their perception of vowel contrasts in analogous fashion to what

was observed in Chapter 2, we can test if glottalization indeed cues prominence to listeners, and if so, how this information is processed online.

The way in which glottalization is implemented in this chapter is purely contextual, i.e. preceding, and not co-occurring with, target vowel information in the stimuli, as in [ʔV]. Here a terminological note is pertinent. As noted above, Garellek (2013) defines a glottal stop as an abstract articulatory target which may be realized in various ways, i.e. a full sustained stop closure made at the larynx, or simply as laryngealized voice quality. The manipulation used in the present chapter is the former, a stop consisting of complete and sustained closure. The term "glottal stop" as used in this chapter accordingly refers to a full glottal stop [ʔ], and not to a more abstract articulatory target.

The presence of a glottal stop is hypothesized to function as a prominence-lending perceptual cue, for the reasons outlined above. In this sense, the manipulation in the experiments in this chapter could be seen as analogous to that in Chapter 2, i.e. a contextual manifestation of prominence. At the same time, prominence cued by a glottal stop seems conceptually different from the prominence manipulation in Experiments 1 and 2. The phrasal prominence manipulation in these previous experiments was intended to convey a change in phonological prominence (accentedness), though phonetic prominence clearly varied across conditions as well. A glottal stop differs from this manipulation in several ways. First, a glottal stop constitutes a single articulatory target, manifested by the adduction of the vocal folds preceding phonation for a following vowel. This is clearly different from phonological prosodic organization which is assumed in many models to involve considerable look-ahead in the speech production process taking long spans of planned speech into account (Keating & Shattuck-Hufnagel, 2002; Krivokapić, 2012, 2014). As a perceptual analog in the prosodic analysis model of Cho et al. (2007), abstract phrasal prosodic organization is computed via prosodic analysis, but localized prominence cues, such as a glottal stop, though they would contribute to a parsed prosodic structure, would not constitute an abstract prosodic representation in their own right. In other words, in keeping with the view of prosodic structure laid out in Chapter 1, [ʔ] does not constitute a prosodic category, but rather manifests as the phonetic encoding of a more abstract, phonological, prosodic structure (e.g.,Keating, 2006; Keating

& Shattuck-Hufnagel, 2002). In this sense, glottalization differs from phrasal prominence in (1) being local to a vowel target and (2) not constituting an abstract (phonological) prosodic category. In this same vein, glottalization varies within phonological prominence categories (e.g., nuclear accented syllables). Dilley et al. (1996) find that though prominent syllables tend to be glottalized, they are not always. If we thus conceptualize glottalization as a phonetic cue to prominence, which can vary within phonological prominence categories, we might predict a different timecourse for its influence, as compared to the phrasal prominence effect in Chapter 2. Recall the timecourse predictions from Chapter 2, which are restated below, with an additional prediction added based on the results in Chapter 2.

(1) *Prosodic analysis* : Formant cues activate lexical hypotheses, and prominence information (glottalization) subsequently modulates lexical competition. The influence of formant information in processing is asynchronous with the influence of glottalization, preceding it in time (e.g., Cho et al., 2007; Kim, Mitterer, & Cho, 2018; Mitterer et al., 2019).

(2) *Phonetic context* : Prominence information (glottalization) is immediately integrated with formant cues, showing a simultaneous influence (e.g., McMurray & Jongman, 2011; Toscano & McMurray, 2015).

(3) *Multi-stage influence* : An overall delayed effect of prominence, but evidence of a weaker early influence in fine-tuning formant perception, analogous to what was observed in Experiment 2.

A contextual prominence cue that does not directly implicate more abstract prosodic processing might be predicted to align with prediction (2), that is, immediate integration of formants and prominence resulting in a simultaneous or near-simultaneous influence, and similar trajectories for each effect, analogous to compensation for preceding rate and spectral context (Maslowski et al., 2019; Reinisch & Sjerps, 2013). We might alternatively predict that the effect of glottalization would show a multi-stage influence, that is, early phonetic effects of prominence, but an overall delayed timecourse, as in Experiment 2. However,

this seems unlikely given that prominence information conveyed by only [ʔ] is not seen as constituting phonological prosodic structure, as discussed above. As such, outcome (2) is predicted in this case. As in Chapter 2, these hypotheses will be tested in two ways: first, in observing overall divergence times obtained from GAMM difference smooths, and second, from inspecting surface plots showing the influence of the continuum across prominence conditions.

The answers to these questions will help us better understand if glottalization serves as a prominence cue and how it is integrated with formants in vowel perception. This more generally will help us explain how localized prominence cues are processed, especially in comparison to phrasal prominence (in Experiment 2).

3.3 Experiment 3

Experiment 3 addressed the questions outlined above by testing if listeners adjust categorization of an /ɛ/-/æ/ ("ebb"-"ab") continuum as they did in Experiment 1, but now with glottalization as a prominence cue. The crucial manipulation in Experiment 3 was accordingly whether or not a glottal stop preceded the target vowel in a vowel hiatus environment where the target is the second vowel in a VV sequence. Following the results of Experiments 1 and 2, it is predicted that a glottal stop, if it cues prominence to listeners, should shift categorization of the continuum in line with sonority expansion effects. More concretely: a glottal stop (as compared to no glottal stop) should show increased "ebb" responses, analogous to the prominence effect in Experiment 1.

3.3.1 Materials

The materials used in Experiment 3 were created by re-synthesizing the speech of a ToBI-trained American English speaker. The speech material was recorded in a sound-attenuated booth in the UCLA Phonetics Lab, using an SM10A Shure^{TM} microphone and headset. Recordings were digitized at 32 bit with a 44.1 kHz sampling rate. The method of stimulus manipulation was the same as that in Experiment 1, implementing Burg method LPC

resynthesis in Praat using a script (Boersma & Weenink, 2020; Winn, 2016). Stimuli from Experiment 3 will also be used in Experiment 4 which will offer a timecourse assessment of listeners' processing of the glottal stop and will be compared to Experiment 2.

The starting point for stimulus creation was a production of "say the ebb now", with "the" produced as [ðə]. This was produced with pitch accents on the word "say" and the target, such that the target bore the nuclear accent. The creation of the continuum only altered F1 and F2 in the target word creating a [ðəɛb] to [ðəæb] continuum, with continuous formant transitions from the precursor vowel to the target. This constitutes what will subsequently be referred to as the "no glottal stop condition", where no glottal stop preceded the target sound in the hiatus environment, as shown in Figure 3.1. The average F1 and F2 values for the stimuli were set to be the same as the endpoints for Experiment 1, to make stimuli in these two experiments more comparable. This entailed a slight adjustment of the [ðəɛb] endpoint of the continuum. The formants in the precursor vowel were also slightly centralized (F1 raised, F2 lowered) to ensure they predicted the opposite of spectral contrast effects (described below). This manipulation made the precursor vowel sound slightly lower than a canonical [ə], though it was still perfectly intelligible and judged to sound natural.

The goal in creating the "glottal stop condition" was to cross-splice [ʔ] from a production of the carrier phrase in which it preceded the target. The portion of the glottal stop that was inserted preceding the target was the silent closure (approximately 100 ms in duration), and the short aperiodic burst that accompanied the release of the stop (approximately 15 ms). Because it was predicted that a preceding glottal stop should increase listeners' "ebb" responses, the production from which [ʔ] was cross-spliced was [ðəʔæb]. This ensures that, in the case that any information about the following vowel is contained in the release of the stop (though none was perceived), it would bias listeners towards /æ/ when a glottal stop precedes the target, which is the opposite of the predicted prominence effect. The point at which the glottal stop was inserted was at the end of the precursor, where formant trajectories began to shift to the target vowel. The insertion of [ʔ] resulted in a sudden end to the vowel in the precursor. To render the precursor more natural, several periods from [ə] in the production of [ðəʔæb] were cross-spliced and appended to the end of the precursor vowel. The endpoint

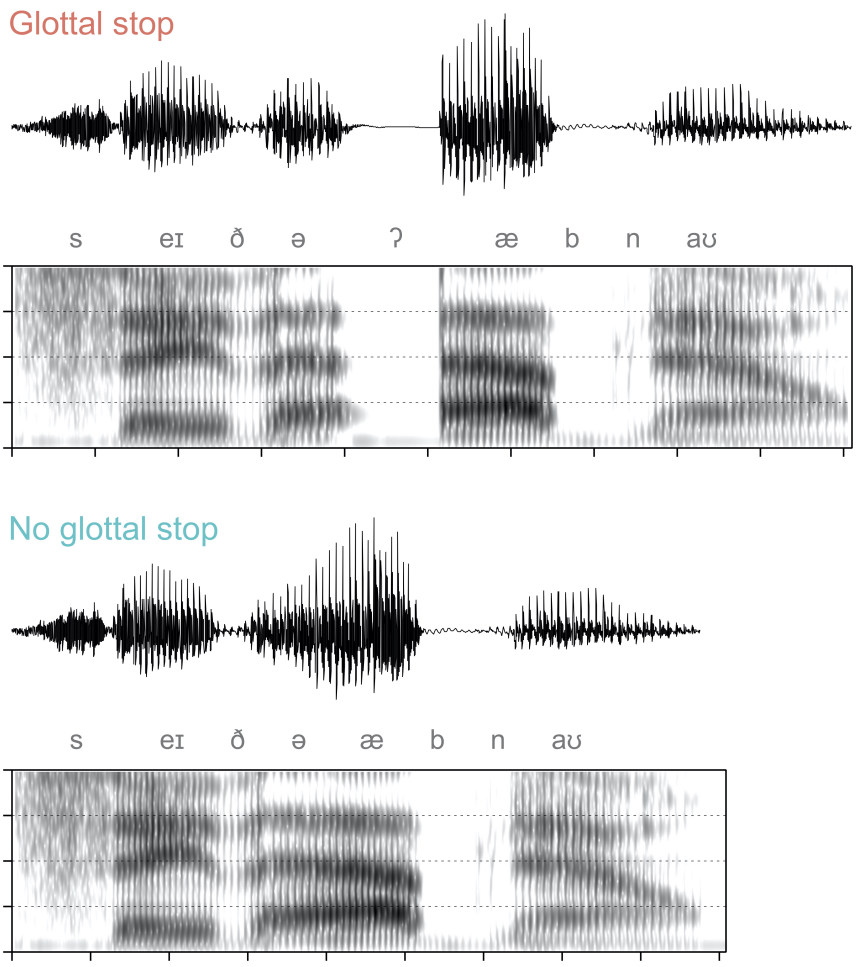

Figure 3.1: Waveforms and spectrograms of the Experiment 3 stimuli. A segmental transcription is given in IPA above the spectrograms. Ticks on the y axis indicate 1000Hz, for a total frequency range of 0-4000Hz. Ticks on the x axis are placed at every 100 ms. The target word shown in the figure is from the /æ/ endpoint of the continuum.

of the precursor vowel therefore showed a dip in amplitude and irregular voicing going into the glottal stop, which improved the naturalness of the stimuli substantially. This modified precursor vowel and following [ʔ] were cross-spliced to precede all steps on the continuum, resulting in a [ðəʔɛb] to [ðəʔæb] continuum, as shown in Figure 3.1. Of note, the target clearly bears a nuclear pitch accent in both conditions (unlike Experiments 1 and 2). This can be seen in the waveforms and spectrograms in Figure 3.1. As such, we can consider this a case where accentedness (or, phonological prominence) does not vary across conditions, but instead, within a pitch accent category, a phonetic parameter varies to signal a difference in prominence (cf. Bishop et al., 2020; Grice et al., 2017).

With this stimulus structure in mind, another possible perceptual effect (unrelated to prominence) merits consideration. Note that a glottal stop introduces a discontinuity in formant trajectories in the two vowel sequence. In the no glottal stop condition, formants transition continuously from the preceding [ə] to the target vowel, while in the glottal stop condition, an interval of silence interrupts the preceding formants in [ə] (see Figure 3.1). This temporal separation between the target and the precursor in one condition, but not the other, might influence perception of the target sound by the mechanism of *spectral contrast* (e.g.,Holt et al., 2000; Stilp, 2020). This is outlined below.

Related to spectral contrast, it is well established that the articulation of a vowel is impacted by other vowels that are local to it, referred to here as vowel-to-vowel (V-to-V) coarticulation (Cho, 2004; Fletcher, 2004; Öhman, 1966; Recasens, 1984). These coarticulatory influences are evident in both anticipatory and carryover contexts. They occur when a consonant intervenes between the two relevant vowels, i.e. in a VCV context (Beddor, Harnsberger, & Lindemann, 2002), and even with more intervening material (Magen, 1997). Acoustically, these effects are manifested in the formant structure of one vowel becoming more similar to another vowel (Beddor et al., 2002; Cole, Linebaugh, et al., 2010; Öhman, 1966). For example, Beddor et al. (2002) found that F1 and F2 of /i/ and /e/ changed systematically based on the vowel that preceded or followed: an adjacent /i/ raised F2 and lowered F1 (rendering the impacted /i/ and /e/ more /i/-like) as compared to e.g., an adjacent /ɑ/, which lowered F2 and raised F1 (rendering the impacted /i/ and /e/ more

/ɑ/-like).

It has further been demonstrated that listeners compensate perceptually for V-to-V coarticulation (e.g., Beddor et al., 2002). A proposed perceptual mechanism behind these effects is *spectral contrast* (e.g., Holt et al., 2000; Stilp, 2020, cf. Fowler, 2006). This refers to the well-established finding in the literature that frequency distributions in the spectrum are perceived by listeners *relative to their context*. Spectral contrast effects can be induced by neighboring consonants (Mann, 1980), vowels (Beddor et al., 2002), long term average spectra (Ladefoged & Broadbent, 1957; Reinisch & Sjerps, 2013), and non-speech stimuli with properties that mirror spectral distributions in speech (Holt, 2006; Holt et al., 2000; Stilp, Alexander, Kiefte, & Kluender, 2010).[1] Though spectral contrast effects can occur with temporal separation between the target and context (Holt, 2005), temporally more local context generates stronger effects, and temporal distance reduces their strength (Holt, 2005; Stilp, 2018, 2020).

Given this, consider the relevance of V-to-V coarticulation and spectral contrast to the stimuli in Experiment 3. Whatever spectral contrast effects are introduced by the precursor vowel, we would expect them to be reduced in strength by an intervening glottal stop (introducing temporal separation between precursor and target), and to be stronger when no glottal stop is present. First consider the no glottal stop condition. Recall that the precursor vowel [ə] was created to have relatively high F1 and low F2. With continuous formant transitions from the precursor to the target sound, we would expect precursor F1 and F2 to impact perception of the target such that, overall, F1 is perceived as relatively lower (following higher F1 in the precursor). For F2, the precursor value falls more in the middle of the continuum on average, though it is slightly lower than most continuum steps. The relationship between precursor F1 and F2 and the steps on the continuum is shown in

[1]Spectral contrast effects are generally assumed to operate at an early and general auditory level of processing, given their existence with non-speech (Holt, 2006; Stilp et al., 2010), in non-human animals (Lotto, Kluender, & Holt, 1997), and immediate impact on processing (Reinisch & Sjerps, 2013). However, the picture is complicated by data showing some contexts generate spectral assimilation effects (the opposite of contrast) as shown in e.g., Repp (1983). The conditions under which contrast and assimilatory effects occur is an active area of research (Rysling, 2017; Rysling, Jesse, & Kingston, 2019). Language-specific perceptual patterns have also been observed (Beddor et al., 2002).

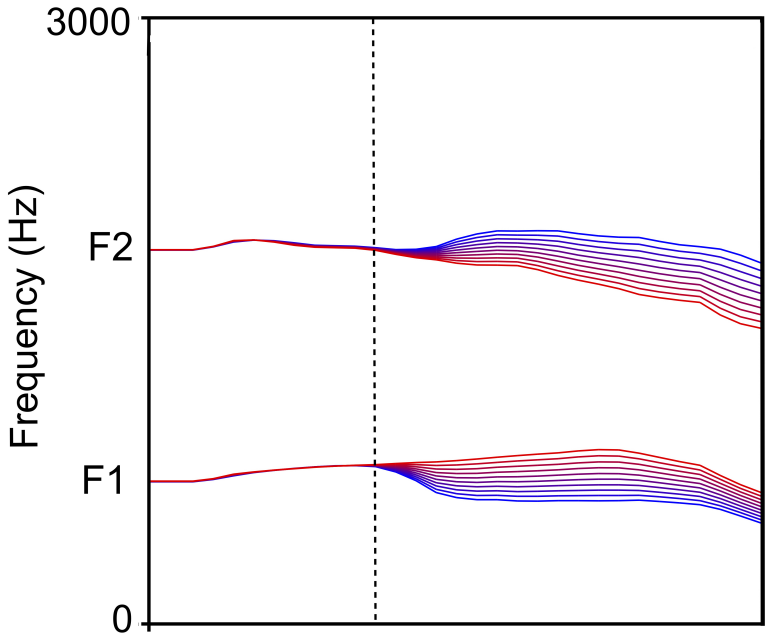

Figure 3.2: Formant tracks showing the Experiment 3 continuum with frequency (0-3000 Hz) on the y axis, and time on the x axis. F1 and F2 are indexed to the left. The point at which [ʔ] was inserted is shown by the dashed vertical line. The continuum is arrayed such that the F1 and F2 values for the /æ/ endpoint are the innermost red lines, and the F1 and F2 values for the /ɛ/ endpoint are the outermost blue lines, as in Figure 2.2.

Figure 3.2.

Higher perceived F1 and (possibly) lower perceived F2 would lead to a more /ɛ/-like perceived target in the no glottal stop condition, as a function of spectral contrast. Compare this to the glottal stop condition. Given that spectral contrast is reduced or eliminated as temporal distance increases between the precursor and target (Coady, Kluender, & Rhode, 2003; Stilp, 2020), and that contrast effects more generally seem to follow this locality constraint (Newman & Sawusch, 1996), we would expect this spectral contrast effect to be reduced such that the target is perceived as relatively /æ/-like in the glottal stop condition. This predicts increased /æ/ responses therein. To put this effect in coarticulatory terms, a glottal stop might reduce perceived coarticulation between two adjacent vowels and there-

fore decrease listeners' attribution of target formant structure to the precursor vowel, i.e. an [əæ] sequence would show a stronger coarticulatory influence on the second vowel, as compared to an [əʔæ] sequence. On this basis, more "ebb" responses would be seen for [əV], as compared to [əʔV]. In sum, spectral contrast effects predict increased "ebb" responses in the no glottal stop condition. This is notably opposite of the predicted prominence effect, where a glottal stop (in the glottal stop condition) should increase "ebb" responses if it cues prominence. This competing prediction rests crucially on the fact that the precursor vowel has more centralized F1 and F2 than the steps of the continuum overall.[2]

One additional consideration is the direct impact of glottalization on perception of vowel quality. It has been remarked that typologically, low vowels and glottalization tend to co-occur, and a perceptual explanation for this pattern is forwarded by Brunner and Zygis (2011). The authors created a continuum from /i/ to /e/ which German listeners categorized as one of two words. The crucial manipulation was whether the vowel was glottalized or not. Notably, the authors manipulated only f0 as a cue to glottalization (e.g., Hillenbrand & Houde, 1996), making their manipulation quite different from the one used here. Listeners in their study also categorized isolated words, unlike the present study. The authors found that a glottalized vowel was perceived as *lower*, that is, showing more /e/ responses overall, and thus conclude that glottalization lowers perceived vowel height.[3] Given that the two vowels used in the present experiment are /ɛ/ and /æ/, an analogous prediction is that glottalization would lead to a lower perceived vowel, increasing "ab" responses in the glottal stop condition. This, like the predictions based on spectral contrast, is the opposite of the predicted prominence effect.

The way in which the stimuli are constructed in Experiment 3 therefore allows for a conservative test for the hypothesis that glottalization cues prominence to listeners. As it pertains to spectral contrast effects, an illustration of why this sort of stimulus design is

[2]Of note, because the precursor vowel is unaccented and fairly reduced, it would be expected to exert more minimal coarticulatory effects on the following accented target vowel in general (e.g., Cho, 2004; Fowler, 1981).

[3]The authors forward this as an explanation for the typological patterning of low vowels and glottalization, where listeners reinterpret glottalized vowels as being lower.

Table 3.1: Model output for Experiment 3.

	Estimate	Est. Error	L-95% CI	U-95%CI	credible?
intercept	1.18	0.16	0.87	1.50	✓
glottal stop	1.75	0.23	1.31	2.22	✓
continuum	-3.35	0.17	-3.71	-3.02	✓
glottal stop:continuum	-0.79	0.20	-1.20	-0.41	✓

necessary is included in Appendix A of this dissertation. The experiment therein, Experiment 7, reports a pilot study in which the use of a different precursor ([i], a high front vowel) led to a confounding influence with the predicted prominence effect. As such it can offer an illustration of the need to consider domain-general contrast effects in stimulus design, in the vein of Mitterer et al. (2016) and Steffman (2019a, 2019b).

3.3.2 Participants and procedure

30 participants were recruited from the same population as previous experiments. The procedure was identical to Experiment 1.

3.3.3 Results and discussion

Results in Experiment 3 were assessed with the same model structure as previous analyses. An "ebb" response was mapped to 1, an "ab" response was mapped to 0. In contrast-coding the glottal stop condition, the glottal stop condition was mapped to 0.5, and the no glottal stop condition was mapped to -0.5. The model output is shown in Table 3.1 and categorization responses are plotted in Figure 3.3.

In a departure from previous results, the model shows that the intercept is credibly different from zero (in log-odds space) (β=1.18, 95%CI =[0.87,1.50]). The positive coefficient suggests an overall "ebb" bias in the continuum, which is apparent in Figure 3.3. This likely originates from the fact the continuum used in Experiment 3 had the same average F1

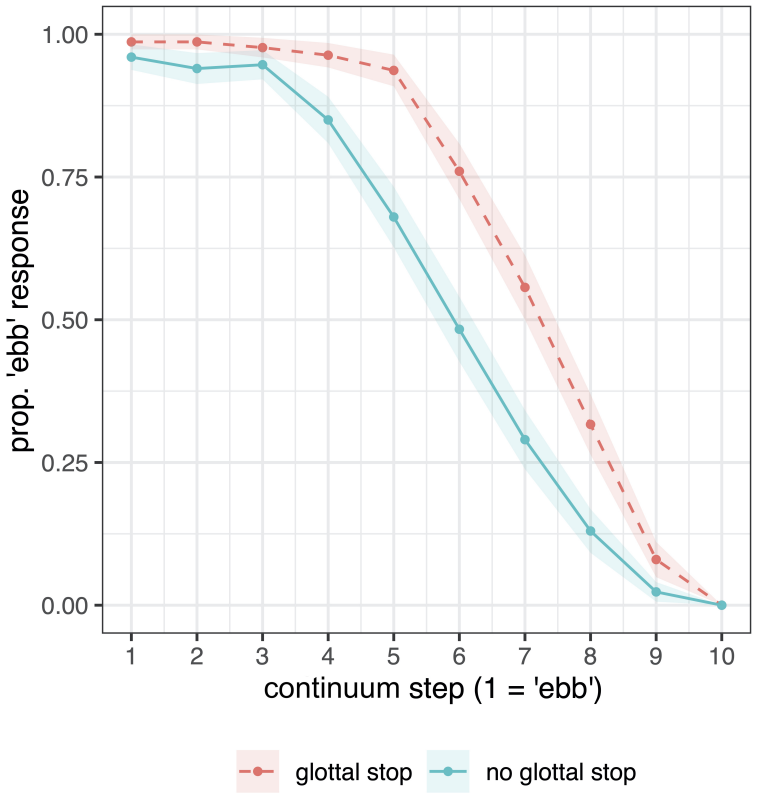

Figure 3.3: Categorization responses in Experiment 3, with the proportion of "ebb" responses plotted on the y axis, split by prominence condition and continuum step, where step 1 is the /ɛ/ endpoint of the continuum. Shading around each line shows 95% CI.

and F2 values as that used in Experiment 1, but was preceded by a different vowel, as compared to the vowel [eɪ] in the word "say" in Experiment 1. Across experiments we could therefore expect a shift in perception of the target via the aforementioned spectral contrast effects (Holt et al., 2000; Reinisch & Sjerps, 2013) such that the relatively lower [ə] used in Experiment 3 would be expected to make F1 and F2 in the target sound more peripheral. That is, target F1 should be perceived as lower (more /ɛ/-like), and target F2 should be perceived as higher (more /ɛ/-like) in Experiment 3 as compared to Experiment 1. The difference in preceding vowel context, with matched average F1 and F2 in the target, is the likely cause of the overall bias towards "ebb" observed in Experiment 3. Nevertheless, categorization at continuum endpoints is strongly anchored (97% "ebb" responses at step

1, 0% responses "ebb" at step 10). Continuum step was also observed to credibly impact categorization in the expected way (β=-3.35, 95%CI =[-3.71, -3.02]). The central predictor of interest, the presence/absence of a preceding glottal stop, was also observed to credibly impact responses, whereby "ebb" responses increased in the glottal stop condition (β=1.75, 95%CI =[1.31, 2.22]). This suggests that in spite of contrast effects that might predict the opposite shift in categorization, a preceding glottal stop shifted listeners' perception of the target vowel in line with the predicted prominence effects (analogous to the effect seen in Experiment 1). This is discussed further below.

Also not observed in previous experiments, a credible interaction between glottal stop condition and the continuum was found (β=-0.79, 95%CI =[-1.20, -0.40]). To further inspect the interaction, contrasts from the model were compared using the package *emmeans* (Lenth, Singmann, Love, Buerkner, & Herve, 2018). The *emtrends* function was used to test for asymmetries in the effect of changing continuum step (a continuous predictor) in each glottal stop condition, providing estimates and credible intervals for the effect of the continuum in each condition. This comparison finds that changing F1 and F2 along the continuum exerts a larger influence in the glottal stop condition (β=-3.73, 95%CI =[-4.16, -3.32]), as compared to the no glottal stop condition (β=-2.95, 95%CI =[-3.28, -2.61]). This asymmetry may be related to the observed "ebb" bias in the continuum. In the glottal stop condition, responses are anchored at the lower steps of the continuum, and then decrease suddenly starting around step 5. In contrast, in the no glottal stop condition there is a more gradual decrease in "ebb" responses along the continuum as continuum step increases, and slightly less anchored responses at lower steps. The interaction therefore can be taken to reflect a difference in the effect of continuum step across conditions. We could take this to suggest that the prominent glottal stop context led to better discrimination of F1/F2 differences along the continuum generating a stronger impact of continuum step in that condition.

We can consider these results in light of other possible influences on vowel perception in these stimuli discussed in Section 3.3.1. First, because spectral contrast predicts the opposite of this observed effect we can be sure that it is not a possible explanation for the results we see here. Similarly, glottalization as a direct influence on vowel perception predicts that it

should lead to a lower perceived vowel (i.e., increasing "ab" responses, following Brunner & Zygis, 2011). This too is the opposite of what was observed here, where a preceding glottal stop lead to the perception of a *higher* vowel (/ɛ/, as compared to /æ/).[4] Thus, we can instead conclude that listeners are indeed exploiting glottalization as a cue to prominence, and adjusting vowel perception in line with sonority expansion such that "strengthened" (or, more sonorous) formant values are expected following [ʔ]. This results is taken to support the predictions outlined above, and to offer some perceptual evidence for the claim that glottalization in American English serves a prominence-marking function (Dilley et al., 1996; Garellek, 2013, 2014). Additionally, this finding shows that localized cues to prominence seem to exert comparable influences in perception to more global, prosodic-structural manifestations of prominence, as seen in Chapter 2 (at least in offline categorization measures). Further implications of these findings are discussed in Section 3.6.

With the finding that a glottal stop affects listeners' perception of vowels, we are now in a position to test how this information is processed by listeners online. Seeing at what point in time the glottal stop information is used in processing, particularly in comparison to the prominence effects in seen Experiment 2, will help address the questions laid out in Section 3.2.

3.4 Experiment 4

Experiment 4 was a visual world eyetracking experiment, testing the effects observed in Experiment 3. The procedure in Experiment 4 was the same as that used in Experiment 2. In this sense, Experiments 3 and 4 are analogous to Experiments 1 and 2.

[4]Why precisely such a difference between this result and Brunner and Zygis (2011) is observed here is unclear, though it may be attributable to the fact that the test languages were different (English versus German, where by hypothesis glottalization plays a stronger prominence-marking role in English). The glottalized words in that study were also categorized in isolation, which may generate different effects.

3.4.1 Materials

The materials used in Experiment 4 were a subset of those used in Experiment 3, selected by taking the three continuum steps on either side of the 50% crossover point of the categorization function from Experiment 3 (as with Experiments 1 and 2). Steps 4-9 on the continuum were selected by this method, and will be referred to as steps 1-6 in what follows (where step 1 in Experiment 4 refers to step 4 in Experiment 3, and so on).[5]

3.4.2 Participants and procedure

36 participants were recruited for Experiment 4 from the same population as previous experiments. All participants had normal, or corrected-to-normal vision as in Experiment 2. The procedure in Experiment 4 was identical to the procedure in Experiment 2: the reader is referred to Section 2.3.2 for a description of the methodology. The visual display, with orthographic representations of "ebb" and "ab" was the same as in that in Experiment 2. As in Experiment 2, there were four practice trials in which each endpoint from the 6-step continuum was presented in each glottal stop condition once. Following this, there were a total of 96 test trials: each of 12 unique stimuli was presented a total of 8 times, with stimulus presentation completely randomized. The experiment took approximately 20 minutes to complete in total.

3.4.3 Results and discussion

3.4.3.1 Click responses

Listeners' click responses in Experiment 4 were analyzed in the same fashion as previous categorization responses, with the same model structure and variable coding as in Experiment 3. As with Experiment 2, the goal in analyzing click responses was to confirm that listeners are showing offline sensitivity to the intended effects, and to replicate Experiment 3. The

[5]The 6 steps selected for Experiment 4 are shifted one higher numerically on the continuum as compared to Experiment 2, where steps 3-8 from Experiment 1 were used. This is attributable to the "ebb" bias observed in the Experiment 3 continuum, which pushed the overall categorization function rightwards.

Table 3.2: Model output for Experiment 4 click responses.

	Estimate	Est. Error	L-95% CI	U-95%CI	credible?
intercept	1.11	0.17	0.78	1.45	✓
glottal stop	2.66	0.30	2.08	3.28	✓
continuum	-2.89	0.22	-3.33	-2.49	✓
glottal stop:continuum	-0.56	0.23	-1.03	-0.13	✓

model output is shown in Table 3.2, while categorization responses are shown in Figure 3.4.

Analysis of click responses lined up rather directly with the categorization responses in Experiment 3. As in Experiment 3, the intercept was observed to be credibly different from zero, showing an "ebb" bias, as can be seen in Figure 3.4. The bias is smaller than in Experiment 3 as shown by the model estimate for the intercept (β=1.11, 95%CI =[0.78, 1.45]). This bias is present in spite of the fact that stimuli were sampled from the ambiguous region of the continuum, suggesting that listeners re-calibrated their perception of the continuum based on the endpoints they were exposed to throughout the trials. Though they were taken from an ambiguous region in the continuum, these endpoints may have been perceived as more oriented towards "ebb", biasing responses overall. Nevertheless, listeners still clearly perceived the continuum as expected, with a credible effect of continuum step (β=-2.89, 95%CI =[-3.33, -2.49]), and fairly anchored responses. As expected, the presence of a glottal stop also had a credible impact on categorization, showing increased "ebb" responses (β=2.66, 95%CI =[2.08, 3.28]). A credible interaction was further observed between glottal stop condition and continuum (β=-0.56, 95%CI =[-1.03, -0.13]), as in Experiment 3. The click responses therefore confirm that listeners in Experiment 4 are showing the same expected sensitivity to both the continuum and the glottal stop manipulation, replicating Experiment 3.

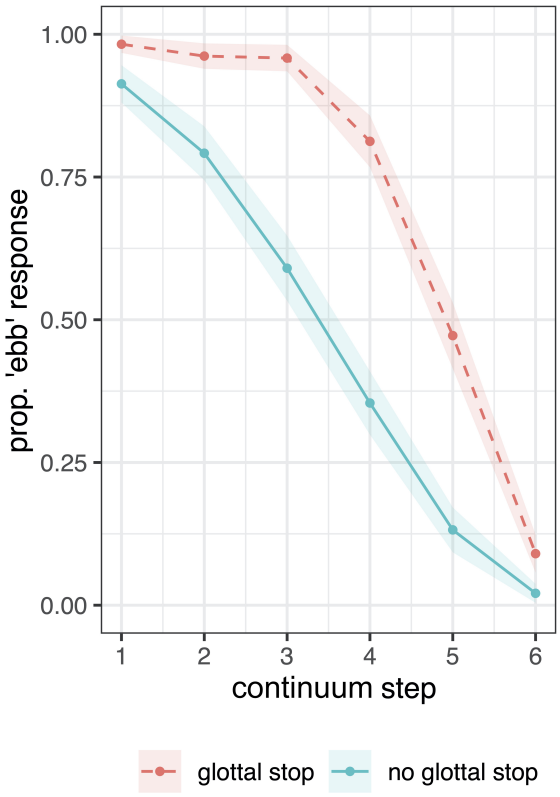

Figure 3.4: Categorization (click) responses in Experiment 4, plotted by prominence condition and continuum step. Note that steps 1-6 in Experiment 4 correspond to steps 4-9 in Experiment 3, as described in the text.

3.4.3.2 Eye movement data

Eye movement data, split by both glottal stop condition and by continuum step, is shown in Figure 3.5. As in Experiment 2, the measure that is visualized in the figure is listeners' "ebb" preference: the proportion of "ab" responses subtracted from the proportion of "ebb" responses at a given time (see Section 2.3.3.2 for details).

Both manipulations had a clear impact on listeners' preference to fixate on the /ɛ/ or /æ/ target. As shown in panel A of Figure 3.5, more "ebb"-like steps generated a preference to fixate on "ebb", with a graded preference for "ab" developing as the continuum steps increase. We can note too that the eye movement data shows an "ebb" bias. Steps 1-3 on

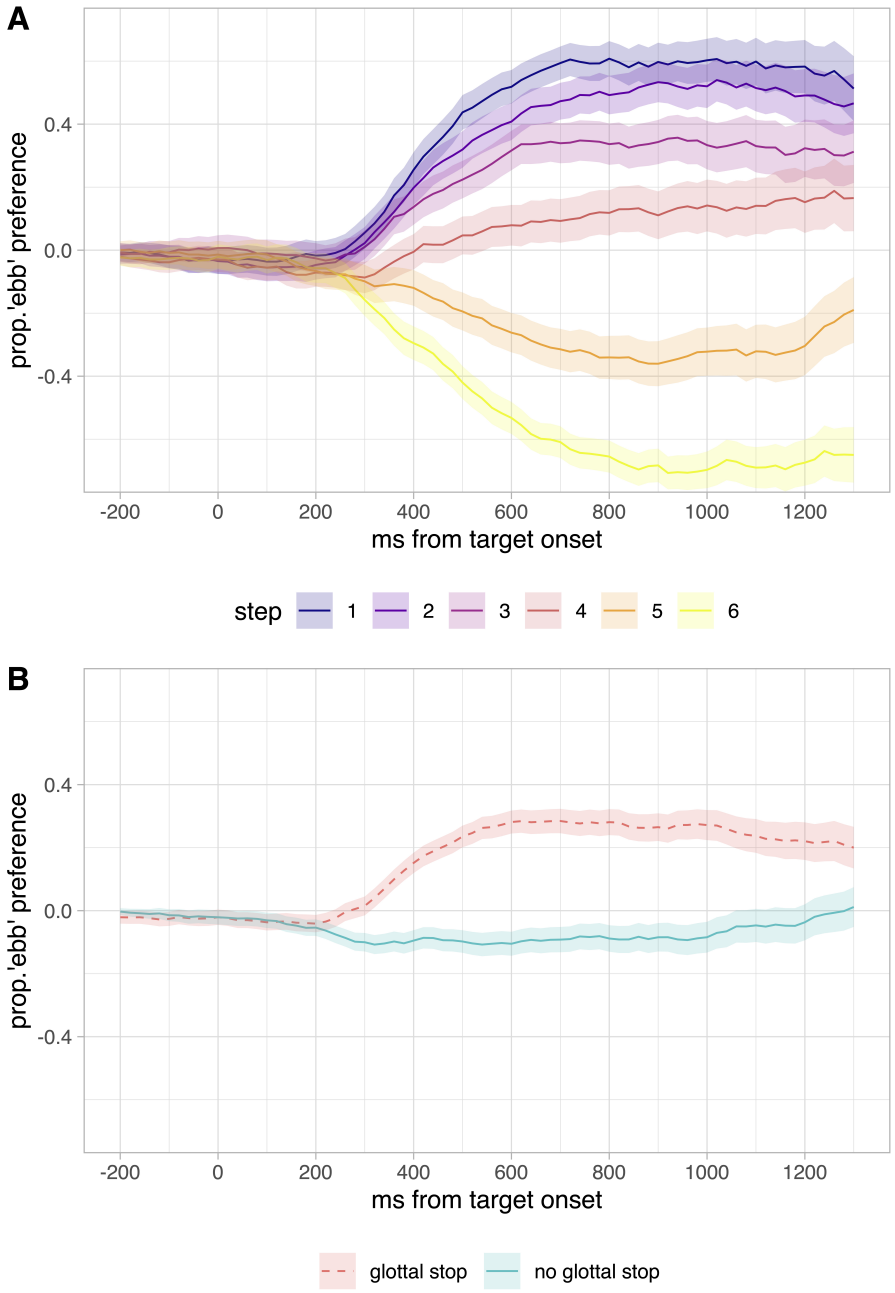

Figure 3.5: Eye movement data for the effect of continuum step (panel A), where step 1 is the /ɛ/ endpoint of the continuum, and prominence manipulation (panel B), in Experiment 4. The x axis shows time ranging from -200 to 1300 milliseconds from the onset of the target sound. The y axis shows the proportion of looks to "ebb" minus the proportion of looks to "ab" (see text). Confidence regions around each line represent 95% confidence intervals, calculated from the raw data.

the continuum show a strong "ebb" preference with step 4 and 5 splitting across zero on the y axis (representing no preference). Nevertheless, a strong "ab" preference (a negative "ebb" preference) is observed for step 6, showing listeners are indeed using the continuum as intended. As shown in panel B of Figure 3.5, the presence of a glottal stop is also exerting a clear influence, where a glottal stop increases looks to "ebb", in line with listeners' click responses.

The statistical assessment of the eye movement data in Experiment 4 was carried out using the same method as in Experiment 2. The GAMM that was fit to model listeners' preference over time had the same model structure as that used in Experiment 2. The reader is referred to Section 2.3.3.2 for a description of the modeling. The numerical model output is given in Table B.2, contained in Appendix B.

The parametric terms in the model confirm the overall patterns seen in Figure 3.5. The intercept shows a clear "ebb" bias, as expected (β=0.37, t = 4.39). The continuum also influenced looks in the overall analysis window, where increasing steps along the continuum decreased listeners' "ebb" preference (β=-1.01, t = -9.86). The glottal stop effect also showed a robust impact on looks in the overall analysis window (β=0.52, t = 4.45).

As with Experiment 3, the timing of the effect of continuum step, and the presence/absence of a glottal stop was assessed by inspecting the divergence between relevant smooths.[6] The effect of continuum step was operationalized as the difference between between steps 3 and 4 on the continuum as in Experiment 2. Note that, though the continuum showed an "ebb" bias as described above, these steps still split across the zero line of Figure 3.5 (showing no preference) early in their divergence.[7] The effect of glottal stop was likewise measured at the scaled continuum value of zero.

As shown by the difference smooth plotted in panel A of Figure 3.6, changes along the continuum exert an early influence on listeners' perception of the target vowel. The estimated

[6]The effect of continuum is shown for smooths in the no glottal stop condition, which differed slightly from the glottal stop condition in being about 30 ms earlier. The estimate that is used therefore represents the *earliest* point in time at which listeners are using information from the continuum, as in Experiment 2.

[7]An alternative would be to measure the difference between steps 4 and 5, which shows a split across the zero line later in processing. This did not differ from the measure derived from steps 3 and 4.

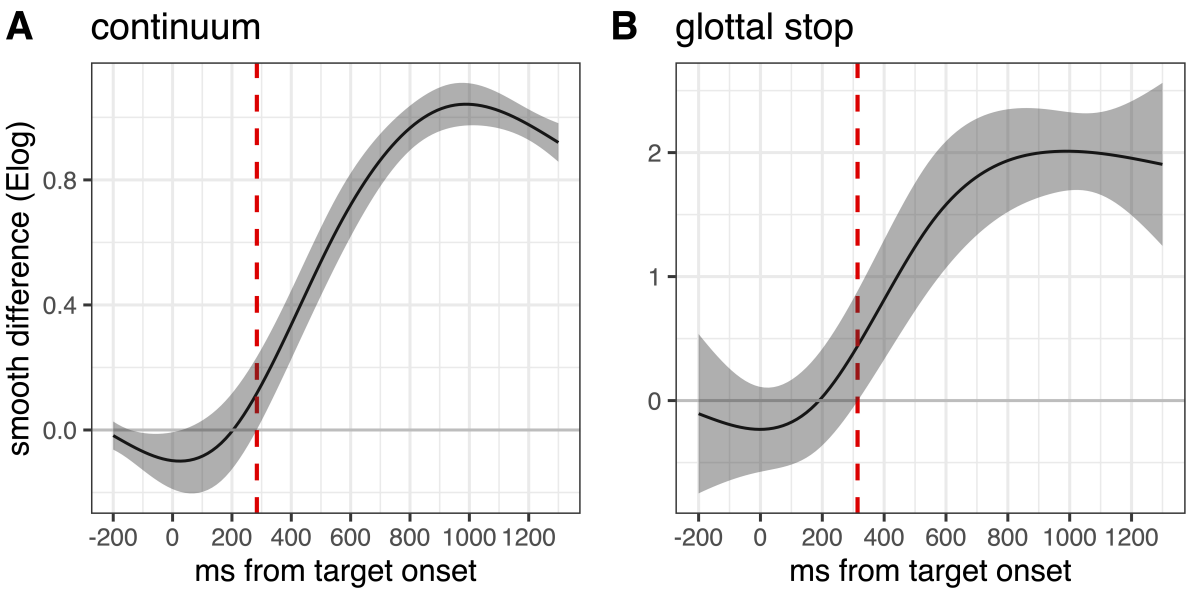

Figure 3.6: Difference smooths (i.e. differences between smooths of interest, as described in the text) for the effect of continuum step (panel A), and the prominence manipulation (panel B) in Experiment 4. The x axis shows time in the analysis window, the y axis shows the difference between smooths in listeners' log-transformed preference measure (see text). Smooths are surrounded by 95% CI, and the red dashed vertical lines index when in time CIs exclude zero, that is, when the difference between smooths becomes reliable (i.e. reliably non-zero). Note the y axes are different in each panel.

divergence time is 284 ms after target onset. This reflects an early effect of the spectral information on lexical activation, lining up with Experiment 2 (for which the continuum effect was estimated at 270 ms), and the general expectation that formant cues should be used early in processing (Reinisch & Sjerps, 2013).

The effect of the glottal stop manipulation evidences an early timecourse as well, with divergence occurring 315 ms after target onset, and following the effect of continuum by approximately 30 ms.[8] This timing relationship is near-synchronous, and can be compared to similar simultaneous influences of context and an intrinsic cue (Maslowski et al., 2020;

[8]In a moving window analysis with 100 ms bins, both the effect of continuum and glottal stop were found to become significant in the 300-400 ms window.

Reinisch & Sjerps, 2013; Toscano & McMurray, 2015). The timecourse shown here, differing by only 30 ms, represents an essentially simultaneous impact in processing, lining up rather straightforwardly with the phonetic context account sketched above (i.e., not reflecting prosodic analysis). The alternative metric reported in Chapter 2, whereby the divergence based on prominence at each continuum step was averaged, yielded only a slightly later estimation of divergence at 345 ms following the target vowel, though notably a glottal stop did not have a significant effect on looks at either continuum endpoint (steps 1 and 6). This is unlike Experiment 2, and is attributable to the fact that listeners' categorization in Experiment 4 is quite anchored at continuum endpoints (see Figure 3.4, cf. Figure 2.4), such that stimuli were essentially unambiguous and context was unimpactful.

This observed early, and near-synchronous, effect means that we should expect to see that the use of formant cues over time varies by glottal stop condition (i.e. the glottal stop shaping listeners' use of formant cues, including early in processing). This was assessed by the same method in Experiment 2: inspecting three-dimensional topographic surface plots in each condition, shown in Figure 3.7. As in Chapter 2, surface plots represent the effect of the continuum over time in each prominence condition, with the shaded region on the plot corresponding to points on the surface where listeners did not have a preference for either target. As with Experiment 2, including glottal stop condition as participating in this non-linear interaction significantly improved the model fit ($\chi^2(5)=1225.28$, $p<0.001$), suggesting an asymmetry across conditions as expected. See Section 2.3.3.2 for more detailed information on interpreting surface plots.

The expected asymmetry across conditions is evident in comparing the panels in Figure 3.7, and noting how the contours of the surface vary. First, in panel A, showing eye movements in the glottal stop condition, we can see evidence of the observed "ebb" bias in the Experiment (i.e. overall more area on the surface being colored with the yellow end of the color scale). Listeners show an immediate and significant "ebb" preference for steps 1-4 on the continuum, with the shaded region shifted strongly downwards. This shows that in the glottal stop condition there is a strong preference for "ebb", and that even very "ab"-like steps in this condition are perceived as ambiguous by listeners (indicated by shading). The rapid

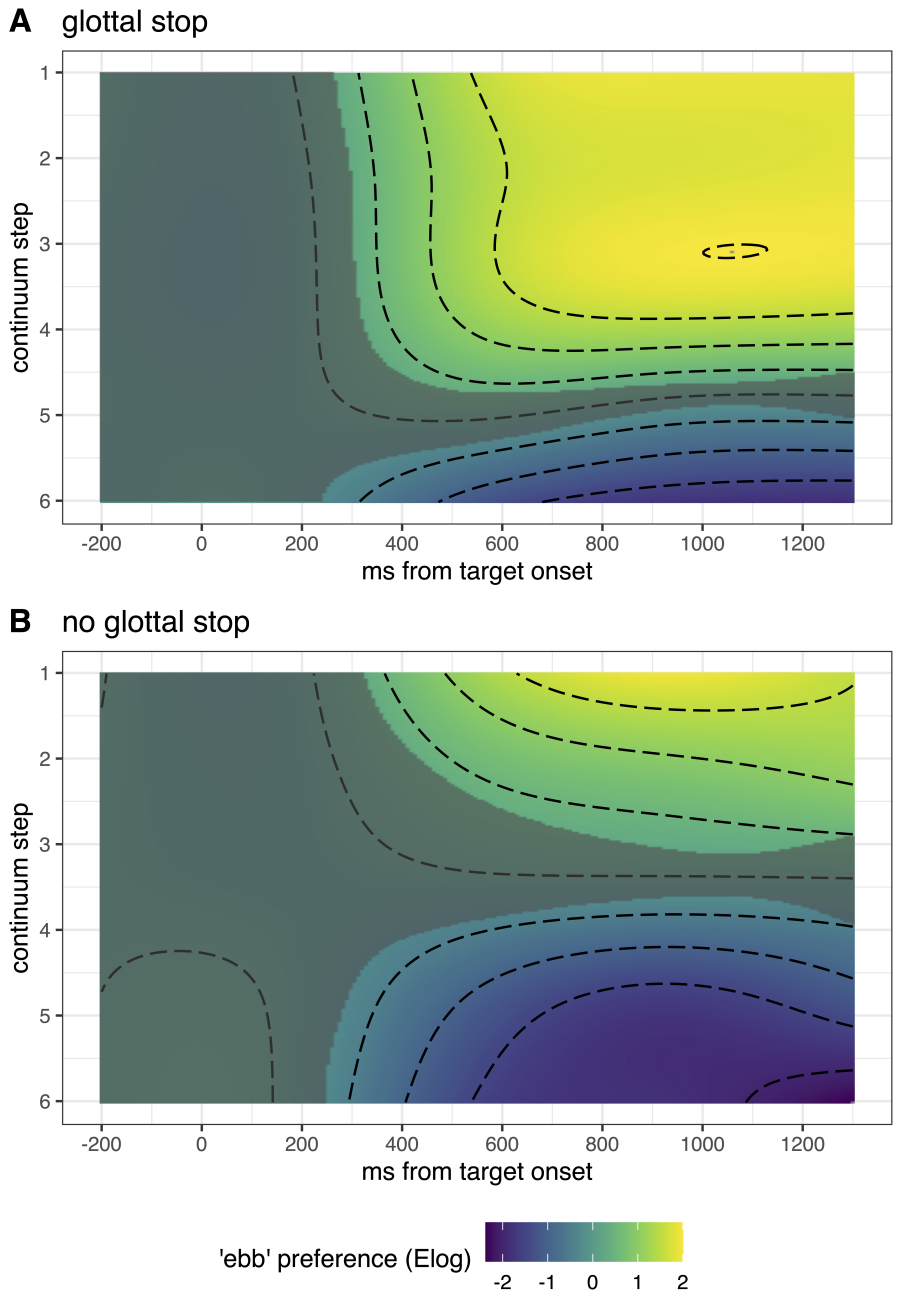

Figure 3.7: Topographic surface plots showing the effect of continuum step (y axis) over time (x axis), split by prominence condition. The color scale represents the degree of listeners' "ebb" preference. Shading on the surface (the darker gray color that covers the leftmost portion of the surface entirely) represents locations in the space for which the preference measure is not significantly different than zero, with 95% CI. Dotted lines show landmarks on the surface.

timecourse of the glottal stop effect is also evidenced by the absence of shading 200-300 ms after target onset, showing an immediate integration of glottal stop and formant information. This can be compared to the no glottal stop condition, in which the shaded region is more centralized in the continuum (i.e. not "ebb"-biased). Though the effect is clearly weaker, we can see that in the no glottal stop condition a stronger "ab" bias develops, as compared to the glottal stop condition, and this preference clearly starts early, at the earliest time that the preference measure becomes significantly different than zero.

The surface plots thus show that, as expected given the early influence of glottalization observed in the divergence measure, listeners' use of formant cues is immediately and strongly impacted by a preceding glottal stop. The shape of the surfaces differs clearly in the earliest time windows, and the regions of ambiguity on the continuum are also shaped by the glottal stop. In other words, the steps on the continuum which are ambiguous to listeners depend heavily on the context. Also relevant is the observation that a glottal stop facilitates looks to a target (particularly to "ebb"), as observed in a smaller shaded area on the surface, which also disappears more quickly over time: 40% of the surface is shaded in the glottal stop condition, while 48% of the surface is shaded in the no glottal stop condition. By separating the target vowel from the preceding context, and by rendering it perceptually more prominent, it appears that a glottal stop aids listeners in recognizing words, offering a further perceptual argument for the prominence strengthening function of glottalization in American English (Garellek, 2013, 2014).

In sum, the surface plots further confirm an immediate and strong influence of preceding glottalization on listeners' perception of the target vowel. This is in agreement with the obtained divergence measures shown in Figure 3.6. Together, these results show an immediate use of the glottal stop which directly shapes listeners' perception of formant cues.

3.5 Comparing Experiment 2 and Experiment 4

As outlined above, one comparison of interest is between the effect of glottalization seen here and the effect of phrasal prominence, observed in Experiment 2. Some evidence for an asym-

Table 3.3: Timecourse summaries for Experiments 2 and 4. "Prominence" refers to the effect of phrasal prominence (Experiment 2), and glottalization (Experiment 4). The difference column shows the difference in the timing of these effects within an experiment.

	continuum	prominence	difference
Experiment 2	270 ms	482 ms	212 ms
Experiment 4	284 ms	315 ms	31 ms

metrical effect can be gleaned just from their timing, as assessed by the divergence between smooths, shown previously in Figure 2.6 and Figure 3.6. Relevant timecourse differences for the effect of phrasal prominence (Experiment 2), and the effect of glottalization (Experiment 4) are summarized in Table 3.3. A glottal stop influenced looks as early as 315 ms after the onset of the target sound, while phrasal prominence showed an effect that followed target onset by 482 ms. By this metric alone we have an indication that these effects reflect different processes (cf. Maslowski et al., 2020). As is clear from the table, prominence effects also vary in their timing with respect to the effect of continuum step, where a substantial asynchrony exists in Experiment 2, but not Experiment 4.

However, as described in Section 2.3.3.2, the effect of phrasal prominence in Experiment 2 is not simply a later-stage influence, but rather it is an effect which starts early, though only becomes more robust later in processing. As compared to the effect of continuum step in Experiment 2, the effect of phrasal prominence reached its maximum later in processing, and grew in a non-linear fashion, as shown in Figure 2.8. Accordingly, to obtain a more dynamic characterization of the effects from each experiment, the range-normalized differences between smooths (from the divergence analyses) were plotted for the effect of the continuum and prominence manipulations in both Experiment 2 and Experiment 4 (as in Figure 2.8). This offers a way to compare how the effect of both the continuum and prominence manipulation changes over time in each experiment, and is shown in Figure 3.8.

As can be seen in the left panel of Figure 3.8, the effect of continuum in each experiment

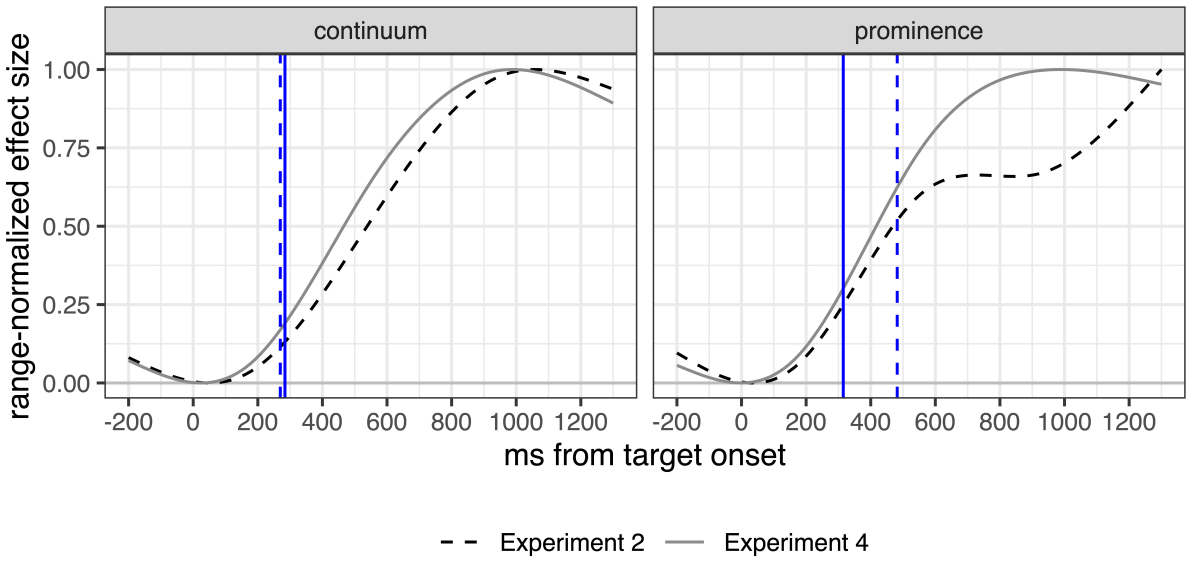

Figure 3.8: Range-normalized effects calculated from difference smooths, for both the effect of continuum (left panel) and prominence (right panel) in Experiments 2 and 4, where a black dashed line corresponds to Experiment 2, and a solid gray line corresponds to Experiment 4 (indexed below the plot). Vertical lines represent when an effect became significant in the divergence-based analysis, corresponding to Table 3.3. Note that "prominence" here refers to prominence manipulations in both experiments: phrasal prominence in Experiment 2, and glottalization in Experiment 4.

is fairly comparable in its trajectory and timecourse. That is, the effect grows steadily in both cases, reaching a maximum at roughly 1000 ms from the onset of the target. This offers some reassurance that, even though the continuum in Experiment 4 showed an "ebb" bias, listeners are using spectral information in generally the same way across experiments. It can also be noted that the effect of continuum grows slightly more quickly in Experiment 4, as compared to Experiment 2, likely driven by the fact that Experiment 4 showed more strongly differentiated responses as a function of continuum step (compare Figure 2.4 to Figure 3.4, and Figure 2.5 to Figure 3.5).

Comparing the effects of prominence in the right panel of Figure 3.8, we can note a clear difference in how the two effects grow over time. The effect of prominence (glottalization)

in Experiment 4 shows a fairly comparable trajectory to the effect of the continuum in both experiments, growing steadily and reaching a maximum slightly before 1000 ms. The effect of phrasal prominence in Experiment 2, as discussed in Chapter 2, is clearly different. This effect reaches a stable state around 600 ms, and then resumes increasing to its maximum at the very end of the analysis window. As such, the maximum effect of glottalization is clearly earlier than that of phrasal prominence.

Experiments 2 and 4 used exactly the same procedure and have the same number of participants. As such, it is hypothesized that these different trajectories reflect different processing mechanisms. A glottal stop shows a robust early effect in processing, integrated essentially simultaneously with formant cues. This reflects an immediate compensatory adjustment in perception, which shows the same general timecourse pattern as e.g., perceptual adjustment for preceding speech rate, or spectral context (Maslowski et al., 2020; Reinisch & Sjerps, 2013; Toscano & McMurray, 2015). This result offers new evidence that prominence information conveyed by a glottal stop is integrated rapidly in processing, a point discussed further in Section 3.6 below. This effect is clearly different from what we see in Experiment 2. The trajectory of the effect in Experiment 2 was taken to reflect a multi-stage influence of prominence, one that fine-tunes early formant processing (via phonetic prominence), but is reinforced later as it is integrated into a more global, phonological, prosodic structure. In comparing Experiment 2 and 4, we can add further nuance to this view: prominence processing needn't show the overall delayed timecourse seen in Experiment 2. Instead, localized prominence cues can trigger immediate compensation. This highlights again how the multidimensional nature of prominence in speech can exert different impacts on processing depending on the nature of the prominence-lending context.

3.6 General discussion

The experiments in this chapter show that the presence of a glottal stop preceding a vowel shifted listeners' perception of formant cues. Experiment 3 tested this claim by constructing a spectral context such that it predicted the opposite of the prominence account, and found

that the prominence effect emerged. This is taken as clear evidence for glottalization as a prominence cue in American English. Given that changes in voice quality, which needn't occur with a full glottal stop, also mark prominence (Garellek, 2013, 2014), further work will benefit from testing how other prominence-driven voice quality features mediate listeners' perception of segmental contrasts (cf. Brunner & Zygis, 2011). The present results additionally predict that we should see analogous influences for other manifestations of prominence strengthening. For example, consonantal prominence strengthening like increased nasal duration in a phrasally prominent NV sequence (Cho & Keating, 2009; Cho et al., 2017), or increased VOT in a C^hV sequence (Cho & Keating, 2009), might be expected to generate the same shifts in perception for formants in a following vowel. Extending the present results along these lines will help generalize the observed effect of glottalization seen here.

Because glottalization is so prevalent at prosodic boundaries in American English (Dilley et al., 1996), it is hard to completely disentangle its boundary-marking and prominence-marking functions. However, following the arguments in Garellek (2013, 2014), the patterning of glottalization with phrasal boundaries could be said to be precisely *because* of prominence marking (counteracting phrase-initial breathy voice quality in vowels) as described in Section 3.2. Nevertheless, glottalization still co-occurs with prosodic boundaries, and as such, further tests of other segmental strengthening cues as outlined above would help offer confirmation that the effects observed here are the direct result of prominence perception. In other words, if various segmental prominence strengthening patterns show the same perceptual result, we would have converging evidence that glottalization too is serving as a prominence cue in this domain. This would be particularly informative in comparing these results to the case of nasal duration in an NV sequence, which shows opposite patterns for prominence and boundary marking as discussed in Section 1.3.2 (Cho et al., 2017).[9]

Further, as observed by Dilley et al. (1996), glottalization and prosodic prominence at the level of the phrase often co-occur. Experiments 1-4 have tested them separately, establishing

[9]Similar converging evidence in favor of this idea could come from testing the influence of glottalization in perception of a high vowel contrast, which shows different acoustic modulations under prominence (discussed in Chapter 4).

that they exert independent effects. However, going forwards, future work might benefit from testing their combined influence in an experiment where they are manipulated orthogonally. Exploring the extent to which their effects are additive, and the relative importance of each, would be a useful extension of the present results.

Experiment 4 explored how glottalization is processed online, and found a robust early influence, nearly simultaneous with the uptake of formant cues, showing an analogous timecourse to compensation for e.g., preceding speech rate, or spectral context (Maslowski et al., 2019; Reinisch & Sjerps, 2013; Toscano & McMurray, 2015). As would be expected given this timecourse, glottalization was shown to have a direct, and early, impact on formant perception, as shown in Figure 3.7.

If, following the discussion above, [ʔ] is taken not to represent a prosodic category per se, but rather to serve as encoding for more abstract prosodic information (Cho et al., 2007; Keating, 2006), the present results show that localized (or, segmental) cues to prominence can modulate perception of vowels independently. This effect is broadly the same as what we see for phrasal prominence (i.e., more sonorant F1 and F2 expected for a prominent vowel). The timecourse data, however, demonstrates that this local cue directly shapes formant perception, and shows the same general trajectory over time as formant information (see Figures 3.6 and 3.7), suggesting a strictly pre-lexical effect. This highlights the multidimensional nature of prosodic prominence, where relevant perceptual cues to prominence include localized modulations such as glottalization.

These timecourse findings thus further show that prominence enters into the multiple stages of processing, and can, in the absence of more global changes in prosodic structure, impact processing immediately, with the same general timecourse as context effects related to speech rate and spectral structure (Maslowski et al., 2019; Reinisch & Sjerps, 2013; Toscano & McMurray, 2015). This differs clearly from the phrase-level (phonological) prominence manipulation in Experiment 2, which showed a slow-growing effect and overall delayed use of prominence in processing (which was preceded by subtle effects of phonetic prominence). A model of prominence in segmental processing and word recognition must therefore allow prominence information to impact multiple stages of processing, and allow access to local

(phonetic/segmental) prominence cues, in the absence of computed phonological prosodic structure (which would generate a delayed influence in processing, unlike Experiment 4). These phonetic cues, when varying *within* a phonological prominence category, as in Experiment 4, exert clear independent and pre-lexical influences in processing.[10] Further implications of these findings for a theory of prosodic, segmental and lexical processing will be discussed and expanded on in Chapter 5.

As a way of further exploring how prominence mediates vowel perception, Chapter 4 extends the idea that phonetic prominence strengthening information is accessed in perception by testing how vowel-intrinsic features mediate prominence effects. This is accomplished by testing how prominence influences perception of a different vowel contrast. The test case, high front vowels, *do not* show consistent sonority expansion (unlike vowels tested in previous chapters), and some studies have reported a conflicting strengthening pattern for them: *hyperarticulation* (see Section 1.3.3). Testing if listeners show sensitivity to vowel-specific prominence strengthening effects will thus better our understanding of how perceptual prominence information interacts with vowel-intrinsic (i.e. featural) properties.

[10]Though phonetic cues will also more generally inform prosodic analysis by contributing, to varying extents, to an overall parsed prosodic structure.

CHAPTER 4

Perceptual prominence effects on high vowels

4.1 The experiments in this chapter

The goal of this chapter is to test how perception of high vowel contrasts is influenced by prominence, exploring if vowel-intrinsic features mediate the effect of contextual phrasal prominence. This chapter accordingly tests how listeners' perception of the American English /i/-/ɪ/ contrast shifts as a function of phrasal prominence. This test case is adopted given that various patterns of prominence strengthening have been documented for the production of high vowels.

Recall that sonority-expanding articulations lead to a more open vowel articulation, i.e. to increased space between the tongue and the roof of the mouth (Cho, 2005; de Jong, 1995). Sonority expansion might thus jeopardize attainment of a high vowel target, and previous studies report that speakers sometimes utilize a different prominence strengthening strategy: *hyperarticulation* (Cho, 2005; Kent & Netsell, 1971). For high front vowels like /i/ and /ɪ/, hyperarticulation would entail lingual fronting and raising, resulting in more peripheral F1 and F2 values (lower F1, higher F2), though notably, previous findings document various different patterns, described below. Experiment 5 accordingly tests whether phrasal prominence will engender perceptual adjustments consistent with hyperarticulation (lower expected F1, higher expected F2) or sonority expansion (higher expected F1, lower expected F2) for these vowels. Experiment 5 further decouples F1 and F2 as acoustic dimensions on a continuum to test if one is more impacted than the other by prominence, a possibility suggested by acoustic and articulatory data in Cho (2005). The results from Experiment 5 are compared to the sonority expansion effects from Experiment 1. Unlike previous experiments,

Experiment 5 was carried out remotely, where participants accessed the experiment via a link and did not come into the lab to complete it.[1] Accordingly, to offer more comparable results to Experiment 5, Experiment 6 was implemented as a remote replication of Experiment 1. This further offers a methodological comparison to Experiment 1, though this is only discussed briefly in this chapter. The results of Experiments 5 and 6 are compared to explore how vowel-intrinsic features mediate phrasal (phonological) prominence effects.

4.2 Conflicting patterns of prominence strengthening

The experiments described in Chapters 2 and 3 showed that listeners incorporate prominence, whether cued phrasally or by a glottal stop, in their perception of vowel contrasts. The contrast under consideration in previous chapters was one for which a clear pattern of prominence strengthening was expected, that is, sonority expansion. However, as discussed in Section 1.3.3, not all vowels undergo prominence strengthening in the same way, and in some cases, strengthening effects appear to be in conflict. In particular, sonority-expanding gestures for high vowels might jeopardize attainment of a high vowel target (Cho, 2005). Some languages, such as Tongan, show uniform raising of F1 in the vowel space for prominent vowels (consistent with sonority expansion), for both high and non-high vowels in the language (Garellek & White, 2015). However, in American English various patterns of prominence strengthening have been documented for high vowels, evidencing both *sonority expansion*, and *hyperarticulation*, as described in Section 1.3.3.

Of particular relevance in this chapter is the case of /i/, tested in various previous studies. Sonority-expanding gestures for /i/ might be detrimental to attainment of the vowel target, perhaps particularly given the necessity of contrast maintenance with /ɪ/ (of note: in Tongan where /i/ undergoes sonority expansion, there is only one other front vowel, /e/). Various prominence strengthening patterns for /i/ have been documented in the literature. For example, Houde (1967), using cineradiographic data for a single speaker, found that prominent articulations of /i/ showed inferior and posterior tongue body displacement, a clear artic-

[1] Due to the novel coronavirus (COVID-19) pandemic.

ulatory manifestation of sonority expansion, similar to that seen for non-high vowels (Cho, 2005; van Summers, 1987). Kent and Netsell (1971), on the other hand (using cinifluorographic data from three speakers) found "[...] tongue-marker positions during stressed /i/ are displaced upward and forward relative to the tongue-marker positions during unstressed /i/" (p 36). This runs counter to Houde's data in showing hyperarticulation effects. Cho (2005), using EMA data from 6 speakers, finds phrasal prominence affects positioning of the tongue such that /i/ shows more extreme articulations in the front/back dimension, that is, lingual fronting. In Cho's data, overall no effect in the vertical dimension emerged, something that could be conceptualized as the "suppression" of sonority-expanding tongue body lowering. These effects translated straightforwardly to the acoustics of /i/ in terms of F1 and F2. Overall no effect of prominence on F1 was found, however F2 showed robust increases in prominent /i/, reflecting fronting, and in feature terms, enhancement of [-back]. Notably too, Cho found that both /i/ and /ɑ/ showed larger lip openings when prominent (though the effects were much larger for /ɑ/). /i/ further did not show robust jaw opening under prominence, unlike /ɑ/. This highlights that different articulatory parameters can encode sonority expansion or hyperarticulation effects for the same vowel articulations (cf. Erickson, 2002).

Also of note in Cho's data is variability in speakers' production of /i/, which highlights possible variation in prominence strengthening strategies. Cho (p 3875) states:

> [...] two speakers showed accent-induced lowering effects in both the acoustic and the articulatory vowel spaces [...]. The tongue lowering for /i/ for these speakers might be interpreted as the entire tongue body being shifted forward along the arc of the palate (as evidenced in the tongue fronting), which may rotate the tongue midposition slightly downward. Alternatively, the lowering of [the tongue] for accented /i/ could be interpreted, not in terms of place feature enhancement, but simply as a byproduct due to the tongue shifting to achieve a proper constriction degree and location in the area of the palate. However, these two alternatives do not fully explain why there is also a corresponding acoustic lowering effect (F1 raising) in the acoustic dimension. Instead, the acoustic and

> articulatory lowering effects observed in some speakers may be interpreted as a result of the articulatory maneuver coupled with the jaw lowering (and the lip opening) to increase sonority.

In another illustration of how these effects on high vowels may be variable, Kim et al. (2016) tested both /i/ and /ɪ/, and measured F1 and F2 for accented and unaccented vowels. This is quite relevant to Experiment 5, where the contrast between /i/ and /ɪ/ is tested. Kim et al. (2016) found that /i/ showed lowering of F1 and raising of F2 under prominence (i.e., hyperarticulation), while /ɪ/ showed some F2 raising (fronting) under prominence, but unlike /i/, did not vary in F1. This result is notable because it suggests that the different vowel categories tested in Experiment 5 might be subject to different prominence strengthening effects. This point will be discussed in Section 4.3.2 below.

Past findings in the speech production literature therefore document various patterns of prominence strengthening for /i/ (and /ɪ/), which we can conceptualize as manifesting two competing influences. On the one hand, syntagmatic contrast enhancement and general increases in phonetic prominence associated with an expanded oral cavity (de Jong et al., 1993; Silverman & Pierrehumbert, 1990). On the other hand, paradigmatic contrast enhancement, strengthening acoustic properties that encode vowel features (de Jong, 1991, 1995). Findings from previous experiments in this dissertation give us a clear expectation that prominence should play a role in the perception of high vowel contrasts, though precisely what the impact should be is less clear, given the various patterns of prominence strengthening attested in the literature.

One other pertinent previous finding comes from perceptual data in Mo et al. (2009). In an RPT task (described in Section 1.3.1), the authors found that changes in formant structure correlated with listeners' perception of prominence in American English speech. Non-high vowels generally showed correlations consistent with sonority expansion, that is, increases in F1 and decreases in F2 were both positively correlated with perceived prominence, indexed by P-scores. For the case of the two relevant high vowels, /i/ and /ɪ/, the pattern was somewhat different. In the F1 dimension, increased F1 correlated with perceived prominence

for both vowels, in line with the pattern evidenced by Houde (1967), and the subset of Cho's (2005) speakers who showed this adjustment in F1. For F2, /ɪ/ showed no correlation with perceived prominence. However, /i/ showed a strong effect whereby increased F2 correlated with perceived prominence, lining up rather directly with Cho (2005). This RPT data might suggest that listeners expect (acoustic) sonority expansion in F1, but hyperarticulation in F2.

The goal of Experiment 5 is accordingly to explore the perceptual effects of prominence strengthening on high vowels. This offers a test of which strengthening patterns listeners prioritize in speech perception, and more generally if different vowel categories undergo different perceptual prominence effects as shown in the speech production literature.

4.3 Experiment 5

Given the possible dissociation for the influence of F1 and F2 suggested by Cho (2005) and Mo et al. (2009), F1 and F2 were varied independently in Experiment 5. The experiment was otherwise quite similar to Experiment 1, with the same phrasal prominence manipulation, implemented in a 2AFC task. The goal will be not only to observe how the high vowels in question pattern with respect to contextual prominence, but also to compare these findings to a remotely implemented replication of Experiment 1 (Experiment 6). In the case that Experiment 5 shows a pattern different from sonority expansion, we would have evidence that perceptual adjustments for prominence strengthening are vowel-specific, that is, vowels are perceived differently not only as a function of contextual prominence, but also that the effect of a prominent context is mediated by vowel-intrinsic properties. This outcome would have a bearing on our conception of the process responsible for these effects, as it would necessitate a more detailed mapping between prominence (prosodic) and vowel (segmental) information in processing. If we adopt the framework of prosodic analysis, it would mean the prosody analyzer needs access to information about how specific segmental categories are strengthened, a point that will be discussed further in light of the results. The extent to which the perception results line up with the various patterns attested in previous articulatory and

acoustic studies will also be discussed.

4.3.1 Materials

A target sound from an /i/-/ɪ/ continuum was placed in the same carrier phrase used in Experiment 1. Listeners categorized this target sound as "seat" or "sit".[2] The carrier phrase was created simply by splicing out the previously used target word in both NPA and post-focus conditions. The goal in doing so was to offer as close a comparison as possible to Experiment 1.

The starting point for creation of the target was the word "sit", produced in the carrier phrase "I'll say sit now" with nuclear prominence as in sentence (1) in Section 2.2.1. This word was excised from the carrier sentence and then set to have pitch and intensity that was the average of a nuclear accented production and a post-focus production of the target in the same carrier sentence, as in sentence (2) in Section 2.2.1. As with Experiment 1, the target was identical across conditions, and prominence was purely contextual.

The target continuum was made by the same method of stimulus manipulation as in previous experiments, altering only F1 and F2. It is worth noting here that a durational difference also exists between /i/ and /ɪ/, where /i/ is longer (e.g., House, 1961; Umeda, 1975). The duration of the target was not manipulated as a cue to the contrast here, and the results in Section 4.3.4 show clearly that F1 and F2 were sufficient to anchor endpoint categorization. The durational difference between /i/ and /ɪ/ will be discussed further in Section 4.3.4 as it pertains to contextual durational contrast effects.

In a departure from previous experiments, F1 and F2 were manipulated independently, such that there were four continuum steps varying orthogonally in each dimension, for a total of 16 steps on the continuum. Steps were evenly Bark-spaced in each dimension, ranging from 3 to 4.8 Bark in F1 space, and 12.4 to 14.2 Bark in F2 space, as shown in Figure 4.1.

[2]These two words were chosen to be relatively matched in frequency, as calculated from the $\text{SUBTLEX}_{\text{US}}$ corpus (Brysbaert & New, 2009). The $\log_{10}$ frequency of "seat" is 3.6, and the $\log_{10}$ frequency of "sit" is 4.2. As previously noted, any frequency bias would be expected to impact overall responses, but not to mediate the effect of prominence.

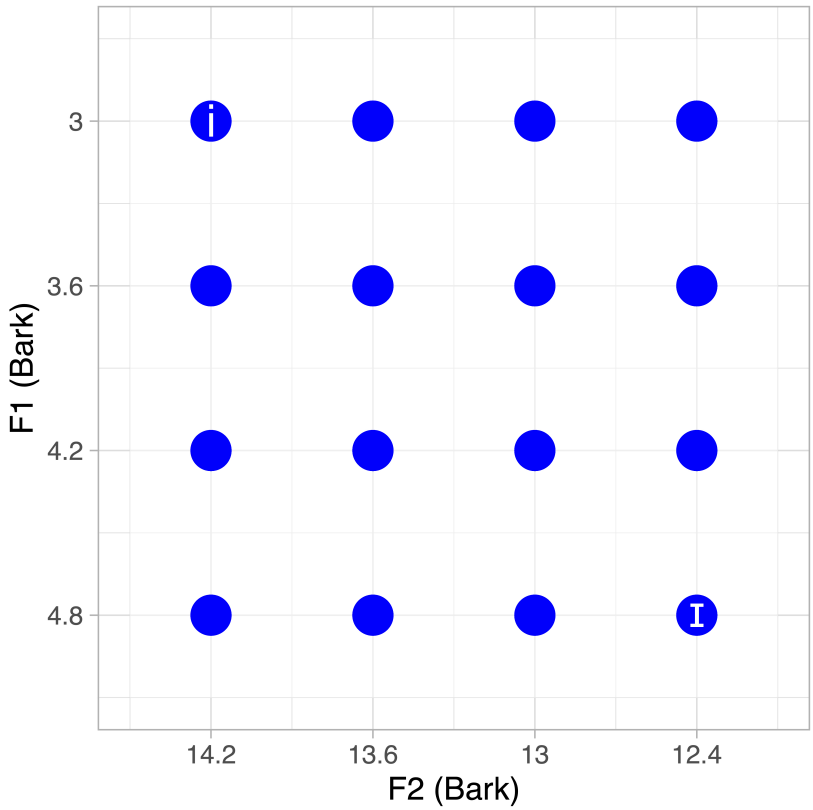

Figure 4.1: A visual representation of the two dimensional continuum used in Experiment 5, with F1 (in Bark) on the y axis, and F2 on the x axis. Note that the axes are reversed to match the typical orientation of the vowel space, with /i/ in the top left corner. Each blue point represents a stimulus token, varying along both dimensions. /i/ and /ɪ/ label the stimuli at the endpoints of the continuum.

Continuum endpoint values were slightly modified from the model speakers' productions of /i/ and /ɪ/, with the F1 dimension expanded slightly to make both dimensions span an equal amount of Barks. The goal in making each dimension equally distributed in Bark space was to ensure that any asymmetrical effects of prominence on F1 and F2 were not due to these dimensions spanning a different amount of perceptual space. Each of the 16 target steps was then spliced back into the carrier phrase frames used in Experiment 1, creating a total of 32 unique stimuli (4 F1 steps $\times$ 4 F2 steps $\times$ 2 frames).

Table 4.1: Predictions for Experiment 5. These predictions apply equally to both F1 and F2 dimensions.

effect	outcome	explanation
sonority expansion	increased /i/ responses when prominent	acoustically less peripheral F1 and F2 values are categorized as /i/ due to expected sonority expansion, such that more tokens are categorized as /i/ overall
hyperarticulation	decreased /i/ responses when prominent	acoustically more peripheral F1 and F2 values are categorized as /i/ due to expected hyperarticulation such that fewer tokens are categorized as /i/ overall

4.3.2 Predictions

Here, given the varying patterns of prominence strengthening observed in the literature, we can contrast two possible outcomes. These are summarized in Table 4.1.

We might also expect F1 and F2 pattern differently as a function of prominence, as discussed above. If it is the case that F1 and F2 undergo different perceptual adjustments under prominence (e.g., sonority expansion in F1 and hyperarticulation in F2 as suggested by Cho, 2005; Mo et al., 2009) then we should expect to see a significant three-way interaction between F1, F2, and prominence in the model. The presence of an interaction could also index a differential prominence effect for the two vowel categories /i/ and /ɪ/, as suggested by Kim et al. (2016). For example, if the /ɪ/-like area of the continuum does not show an effect of prominence in the F1 dimension, but the /i/-like area does (in line with their data, described in Section 4.2), we should expect this to be observable in an interaction between F1 and prominence, where /ɪ/-like F1 is not impacted, but as F1 becomes more /i/-like, prominence exerts an effect.

4.3.3 Participants and procedure

Unlike previous experiments, this experiment was carried out remotely over the internet, though it recruited participants from the same pool as was used in all previous experiments. Online recruitment of participants via platforms such as Amazon Mechanical Turk has been used in tasks collecting acceptability judgments or transcriptions of speech (Marge, Banerjee, & Rudnicky, 2010; Sprouse, 2011), though seemingly less phonetics and speech perception research has been carried out in this fashion. However, recent studies, which have explicitly compared in-lab and remote participant populations, have generally validated the soundness of remote data collection in speech perception experiments (Heffner, Newman, & Idsardi, 2017; Slote & Strand, 2016). Furthermore, the platform that was used to present the stimuli to participants was the same as that used during the in-lab presentation (in Experiments 1, 3 and 4), with the same visual display and instructions.[3]

38 participants were recruited from the same population as in previous experiments (all were students at UCLA and received course credit for participation). Participants were instructed to complete the experiment while wearing headphones in a quiet room. The procedure in Experiment 5 was otherwise identical to other experiments. As with all previous experiments, participants completed four training trials during which they heard each stimulus endpoint (the stimuli labeled /i/ and /ɪ/ in Figure 4.1), in each prominence condition. The test trials consisted of 8 randomized repetitions of each of 32 unique stimuli, for a total of 256 trials (4 F1 steps × 4 F2 steps × 2 prominence conditions × 8 repetitions).

4.3.4 Results and discussion

Results were assessed by the same method as previous categorization data. The model predicted listeners' responses, with "sit" mapped to 0 and "seat" mapped to 1, as a function of F1 and F2 (both scaled and centered as continuous variables), and prominence condition

[3]Of note, the platform Appsobabble (Tehrani, 2020), which was used for all behavioral categorization experiments, runs over the internet on a web browser. The differences between the previous experiments and Experiment 5 were the acoustics of the room in which they were carried out, the headphones and computer used, etc.

Table 4.2: Model output for Experiment 5.

	Estimate	Est. Error	L-95% CI	U-95%CI	credible?
intercept	-0.54	0.15	-0.84	-0.25	✓
prominence	-0.26	0.08	-0.42	-0.10	✓
F1	-1.80	0.15	-2.10	-1.52	✓
F2	2.63	0.18	2.28	2.99	✓
F1:F2	0.78	0.11	0.57	1.00	✓
F1:prominence	-0.01	0.10	-0.19	0.19	
F2:prominence	0.01	0.11	-0.20	0.22	
F1:F2:prominence	-0.01	0.10	-0.20	0.19	

(with NPA mapped to 0.5, and post-focus mapped to -0.5). All interactions were included in the model, and the random effect structure included these three factors and their interactions as by-participant random slopes. The model's fixed effects are shown in Table 4.2.

Firstly, we can note the model intercept is credibly different than zero (β= -0.54, 95%CI =[-0.84,-0.25]), showing an overall "sit" bias. This bias, though robust, is small and endpoint categorization was seen to be well-anchored (discussed below). Turning to the effect of F1 and F2, both stimulus dimensions showed an expected credible effect on categorization. As F1 increased, "seat" responses decreased (β=–1.80, 95%CI =[-2.10,-1.52]). Additionally, as F2 increased, "seat" responses increased (β=2.63, 95%CI =[2.28,2.99]). Both of these outcomes are expected, given that /i/ has lower F1 and higher F2 relative to /ɪ/, and these effects thus show that listeners are using the continuum as intended. The larger effect found for F2 additionally suggests that listeners are more impacted by changes in F2 than F1 on the continuum, i.e. in this case F2 is a stronger cue to the contrast (recall that both dimensions spanned an equal frequency range in Bark). The influence of the continuum on listeners' categorization is shown in Figure 4.2, where listeners' proportion of "seat" responses at each continuum step is given. As is visible in the figure, categorization shifts in a gradient fashion as a function of both F1 and F2, and is well-anchored at the continuum endpoints. At the

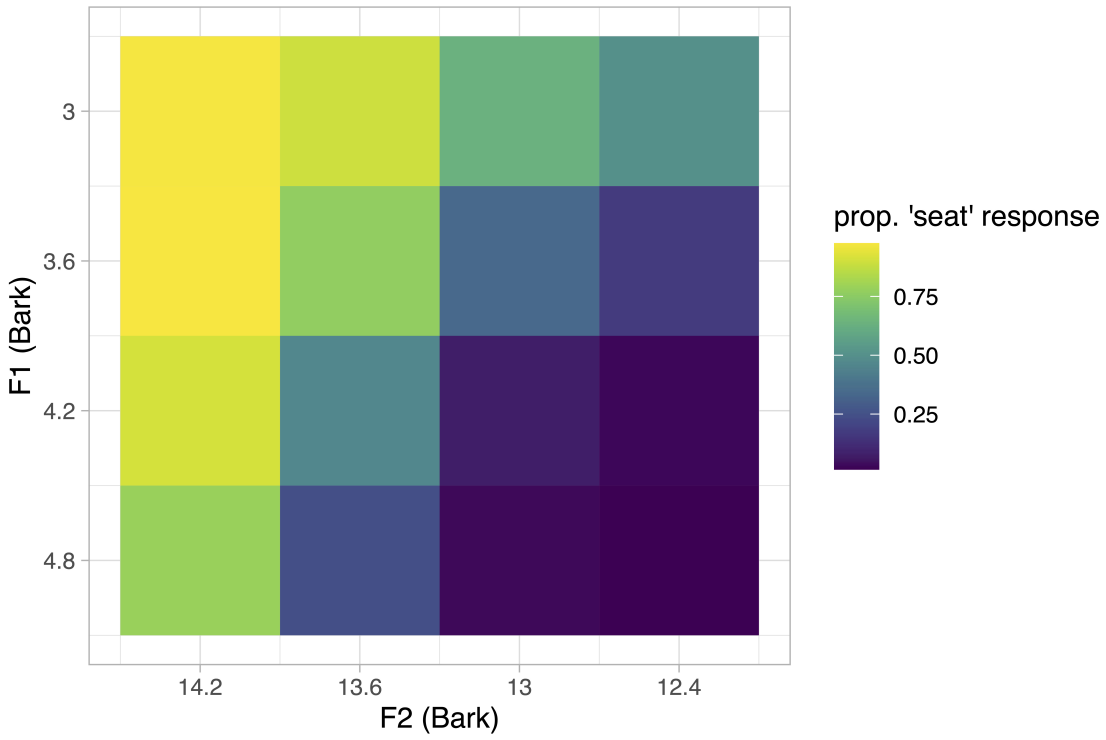

Figure 4.2: Overall categorization responses in the two-dimensional F1/F2 continuum, with F1 on the y axis and F2 on the x axis. Axes are reversed in analogous fashion to Figure 4.1. The color scale at right shows the proportion of "seat" responses at each continuum step.

most /i/-like continuum step, listeners categorized the target as "seat" approximately 97% of the time. At the most /ɪ/-like continuum steps listeners categorized the target as "seat" approximately 1% of the time.

A credible interaction was further found between F1 and F2, suggesting that the effect of one acoustic dimension varied based on the other. Because the interaction of two continuous variables can be difficult to interpret based on the model coefficients alone, the model fit for the effect of F1 (split by F2) and F2 (split by F1) was visualized, and is shown in Figure 4.3. Panel A shows the effect of scaled F1 on listeners' responses at different levels of F2, as estimated by the model. First to note is the general effect of F1: as F1 values increase along the x axis (becoming more /ɪ/-like), "seat" responses decrease, i.e. the lines in panel A are all generally downward sloping from left to right. Next, note how the slope of each line

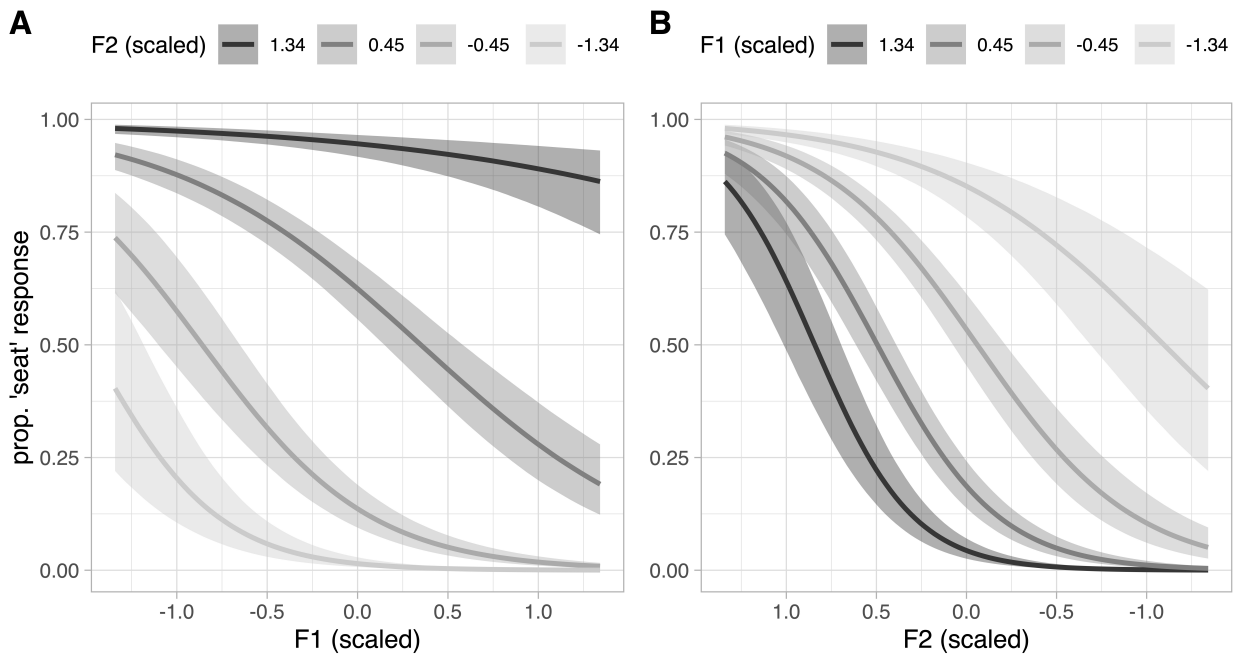

Figure 4.3: Model fit showing the F1 by F2 interaction in Experiment 5. Shading around fit lines shows 95% CI from the model. In the scaled values from the model fit, a *negative* value corresponds to a lower value of F1/F2, a *positive* value corresponds to a higher value of F1/F2. Panel A at left plots scaled F1 on the x axis and represents F2 with lines of varying darkness (indexed above the plot). Panel B plots F2 on the x axis. Note that the x axis in panel B is reversed to match Figures 4.1 and 4.2.

varies. At the three lowest, most /ɪ/-like, scaled values of F2, there is a clear downwards slope, i.e. F1 is impacting responses at these F2 steps. However at the highest F2 step (scaled value 1.34), there is very little impact of F1, that is, the slope of this step's line is very shallow. In panel B of Figure 4.3, the plotting of F1 and F2 is reversed: F2 is on the x axis (ranging from high to low values left to right) and the lines on the plot represent different F1 values. Here an analogous pattern emerges. At higher, more /ɪ/-like, values of F1, F2 has a stronger effect, i.e. the slopes of the fit lines are steeper. As F1 becomes lower (more /i/-like), F2 has a smaller effect, showing shallower slopes across the x axis. What these plots show together is that, for both F1 and F2, at the more /ɪ/-like values of both F1 and F2, the other formant frequency exerts a stronger influence on categorization. The

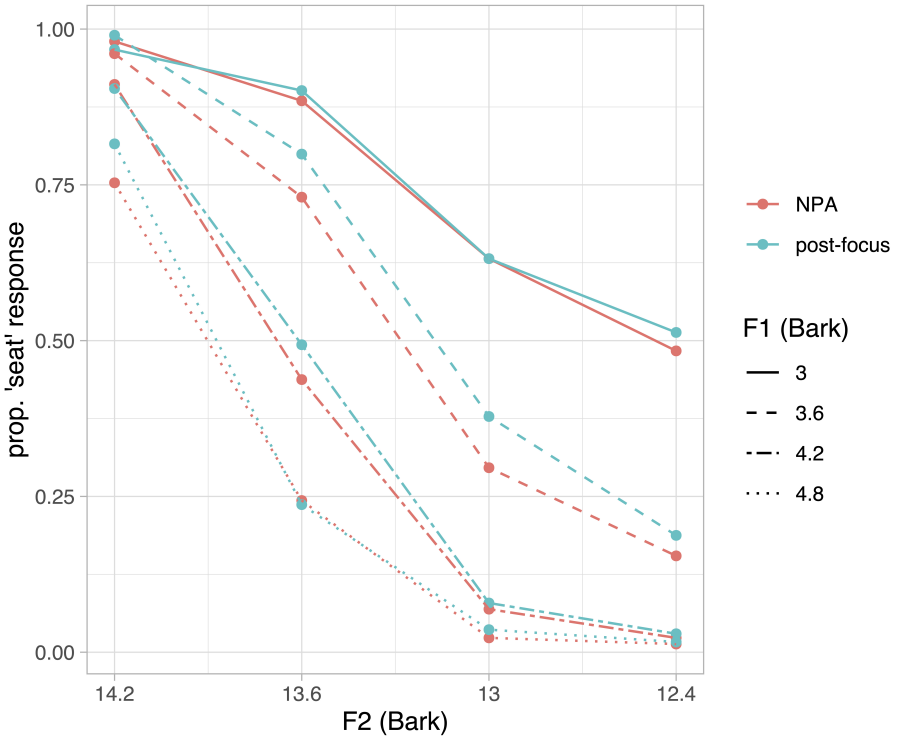

Figure 4.4: Categorization responses in Experiment 5 split by prominence. F2 values are plotted on the x axis, while F1 values are indexed by line type, labeled at right. Note that the x axis is reversed to match Figures 4.1 and 4.2.

interaction can therefore be taken to show that /ɪ/-like spaces of the continuum are more impacted by changing formants, i.e. steps with /ɪ/-like F1 values are shifted more on the basis of F2, and simultaneously, steps with /ɪ/-like F2 values are more impacted by changing F1. On the other hand, when either F1 or F2 signals /i/, the other dimension exerts less of an influence. This is especially true for F2, where the highest F2 step is almost always categorized as /i/, even when F1 becomes quite high as well (see the topmost fit line in panel A of Figure 4.3).

Prominence, the main point of interest in the experiment, also showed a credible effect on responses, whereby listeners showed reliably *decreased "seat" responses* in the prominent NPA condition (β=-0.26, 95%CI =[-0.42,-0.10]). The effect of prominence is visualized Figure 4.4. Note the figure is analogous in orientation to panel B of Figure 4.3, and shows the effect of

prominence such that lines representing categorization in the NPA condition are generally below those in the post-focus condition, though this is more clearly the case at values of F1 and F2 that are in the central region of the continuum. This result shows more peripheral (/i/-like) formant values are expected by listeners in the NPA condition, supporting the hyperarticulation account described above (see Table 4.1). In other words, when the target vowel was phrasally prominent, acoustically more peripheral formants (higher F2 and lower F1) were required for a "seat" response.

As with Experiment 1, in interpreting the prominence effect we can consider domain-general durational contrast effects, given that /i/ is longer than /ɪ/ as mentioned in Section 4.3.1 (e.g., House, 1961; Umeda, 1975). Recall that the vowel /eɪ/ preceding the target is longer in the post-focus condition, as compared to the NPA condition (see Figure 2.1). Durational contrast should make the target vowel sound relatively short (/ɪ/-like) following a longer vowel in the post-focus condition. In comparison, a shorter preceding vowel in the NPA condition would make the target sound relatively long (/i/-like). This predicts increased /i/ responses in the NPA condition, the opposite of the effect we observe. Therefore as with Experiment 1, we can accordingly be sure that domain-general durational contrast is not a possible explanation for the observed effect.

Also important in the interpretation of these results are the effects that are *not* credible. The interactions of F1:prominence and F2:prominence were both observed not to be credible in the model. This suggests that, as in Experiments 1 and 2, perception of the continuum (i.e. the slope of the categorization function for F1 and F2) does not vary across prominence condition. The absence of these credible interactions also speaks against the possibility that listeners are showing different perceptual patterns for the two vowel categories /i/ and /ɪ/, as suggested by data from Kim et al. (2016) and discussed in Section 4.3.2. Additionally, the three-way interaction between F1, F2 and prominence was not credible, nor did it approach credibility (β=-0.01, 95%CI =[-0.20,0.19]). This interaction was predicted to be credible if listeners' perception of F1 or F2 was impacted differently by the prominence manipulation (suggested as a possibility by Cho, 2005; Mo et al., 2009). This outcome thus suggests that perception of both F1 and F2 along the continuum are being impacted in a uniform fashion

by the prominence manipulation, i.e. both showing (acoustic) hyperarticulation effects.

In summary, the results of Experiment 5 show clearly that listeners require more acoustically peripheral (i.e. hyperarticulated) formant values to perceive /i/ in the prominent NPA condition. In comparison to previous findings showing sensitivity to sonority expansion, Experiment 5 shows that vowel-intrinsic features (e.g., vowel height) mediate prominence effects. The total support for the hyperarticulation account offered in these results may be somewhat surprising, given the previously outlined variability in these effects, and the apparent differential relevance of F1 and F2 in this domain (Cho, 2005; Mo et al., 2009). This point is further discussed in Section 4.5, after the variability of the effect in Experiment 5 is investigated and compared to sonority expansion effects.

4.4 Experiment 6: Replicating Experiment 1 remotely

One point of interest given these results is the comparison of the effect found here to what was observed in Experiment 1, where listeners shifted their categorization of the /ɛ/-/æ/ contrast in line with sonority expansion. Recall that Experiment 5 was carried remotely, and Experiment 1 was carried out in a lab setting. To render the results of Experiment 1 more comparable to Experiment 5, a remote replication of Experiment 1, Experiment 6, was carried out. Experiment 6 can additionally serve as the basis for a methodological comparison to the results of Experiment 1, comparing in-lab and remote data collection methodologies, in the same vein as e.g., Heffner et al. (2017). This latter comparison is of only tangential interest in the present chapter, though it will be touched on briefly. The results from Experiment 6 are outlined below, after which they will be compared to Experiment 5.

4.4.1 Materials, participants and procedure

38 participants were recruited online from the same population as previous experiments. The procedure in Experiment 6 was identical to Experiment 1, with the same instructions, number of practice trials, test trials, visual presentation, etc. (see Section 2.3.2). The stimuli in the experiment were the same stimuli used in Experiment 1, categorized as “ebb” or “ab”.

Table 4.3: Model output for Experiment 6.

	Estimate	Est. Error	L-95% CI	U-95%CI	credible?
intercept	-0.14	0.16	-0.44	0.17	
prominence	0.49	0.24	0.01	0.95	✓
continuum	-2.61	0.22	-3.05	-2.18	✓
prominence:continuum	-0.46	0.12	-0.71	-0.25	✓

4.4.2 Results and discussion

Results were assessed statistically by the same method as used for previous categorization data, with the same coding of model variables as Experiment 1. The model output is shown in Table 4.3, and categorization responses are plotted in Figure 4.5.

As seen in Figure 4.5, the prominence effect in Experiment 1 was replicated: the prominent NPA context credibly increased listeners' "ebb" responses (β=0.49, 95%CI =[0.01,0.95]). Notably, the effect is smaller as compared to Experiment 1 (where β=0.83), though the error in each model is comparable (0.28 in Experiment 1 as compared to 0.24 in Experiment 6). The continuum also generally impacted categorization as would be expected (β=-2.61, 95%CI =[-3.05,-2.18]). Unlike Experiment 1, a credible interaction between prominence and continuum was observed (β=-0.46, 95%CI =[-0.71,-0.25]). It can be noted that in Experiment 1, though the interaction was not credible, 95% CI in that experiment only narrowly included zero (β=-0.24, 95%CI =[-0.50,0.01]), and in that sense it is perhaps not surprising that the interaction became credible in Experiment 6. To further inspect the interaction, the *emtrends* function from the package *emmeans* (Lenth et al., 2018) was used to evaluate the effect of continuum step in each prominence condition. This comparison finds a larger effect of changing continuum step in the NPA condition (β=-2.82, 95%CI =[-3.27,-2.37]), as compared to the post-focus condition (β=-2.36, 95%CI =[-2.79,-1.94]). This can be observed visually in Figure 4.5, where we can see that the categorization function in the NPA condition spans more vertical space on the y axis as compared to the post-focus condition. This is

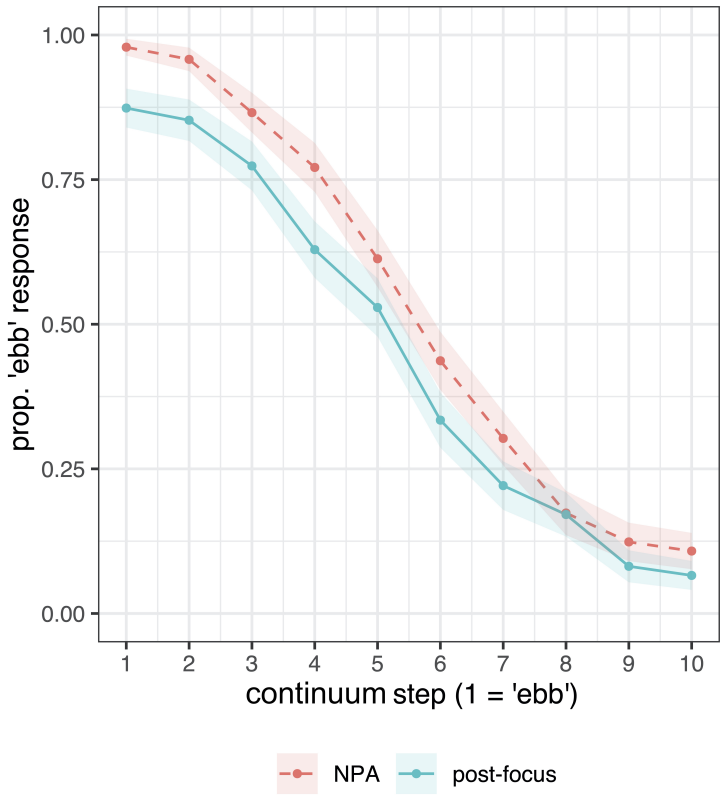

Figure 4.5: Categorization responses in Experiment 6, with the proportion of "ebb" responses plotted on the y axis, split by prominence condition and continuum step, where step 1 is the /ɛ/ endpoint of the continuum. Shading around each line shows 95% CI.

consistent with the idea that listeners better discriminated steps along the continuum when prominent, and lines up with a similar interaction observed in Experiments 3 and 4, where a preceding glottal stop facilitated use of the continuum in terms of categorization responses. These interactions, taken together, suggest that prominence helps listeners discriminate formant differences, showing a larger effect of changing F1 and F2 in prominent contexts (also consistent with the eye movement data from Experiments 2 and 4).

As is also clear in Figure 4.5, separation along the categorization function based on prominence is larger at the lower, more "ebb"-like steps of the continuum. This was further inspected using *emmeans* to test the effect of prominence at each continuum step, as shown in Table 4.4. The comparison shows that indeed, a credible effect is found only at the lower

Table 4.4: The effect of prominence at each continuum step in Experiment 6.

continuum step	1	2	3	4	5	6	7	8	9	10
Estimate	1.21	1.05	0.89	0.73	0.57	0.41	0.25	0.09	-0.07	-0.23
credible?	✓	✓	✓	✓	✓					

steps of the continuum. This might suggest that /ɛ/-like steps on the continuum are more subject to the sonority expansion effect, which could be explained by a difference between /ɛ/ and /æ/ found by Mo et al. (2009) in their RPT study. Perceived prominence was found to be more strongly correlated with F1 and F2 changes in /ɛ/, as compared to /æ/, especially in terms of F2. This could be related to the fact that /ɛ/, being a mid vowel, has more room to "expand" its sonority as compared to the already sonorous /æ/, rendering this location on the continuum more subject to the prominence manipulation. This, in combination with the overall more anchored categorization in the NPA condition (spanning more vertical space on the plot), shows what is driving the observed interaction.

In summary, though there are some differences between the results of Experiment 6 and Experiment 1, the same main effect of prominence is observed in both, serving to replicate this finding with participants who took the experiment remotely. Further comparison of Experiments 1 and 6 is discussed briefly below, though as noted above it is of secondary interest in this chapter. With the results from Experiment 6 in hand, we can now compare them to the hyperarticulation effect in Experiment 5.

4.5 Comparing hyperarticulation and sonority expansion effects

There are two principal points of interest in this comparison. It is clear from simply observing the results of Experiments 5 and 6 that the effect of prominence is different, showing that listeners expected hyperarticulation in Experiment 5 and sonority expansion in Experiment 6. However, less apparent is the magnitude of each effect. One reason to expect a difference in this regard is as follows. Sonority expansion involves robust lowering of the jaw in non-

high vowels (Cho, 2005; de Jong et al., 1993), displacing the tongue body downwards in a relatively large articulatory adjustment. By comparison, lingual hyperarticulation for a vowel like /i/ could be described as a relatively small articulatory modulation.[4] As shown by EMA data in Cho (2005), accented /ɑ/ showed vertical displacement larger than articulatory adjustments associated with hyperarticulation in /i/. Acoustically, F1 and F2 in /ɑ/ showed greater variation overall and as a function of prominence, and more generally, the acoustic consequences of sonority expansion appear to be larger than hyperarticulation in terms of both articulation and acoustics, particularly for lower vowels (see also Babel, 2009; Beckman et al., 1992). Larger shifts in vowel acoustics as a function of sonority expansion would accordingly lead us to predict that perceptual adjustments for this effect may be larger than those for hyperarticulation.

Secondly, we can explore if and to what extent participants vary in how they are impacted by prominence in each experiment. Given that hyperarticulation effects for /i/ appear to be variable according to previous studies, we might expect more variable listener responses, i.e., some participants may favor sonority expansion while others favor hyperarticulation, perhaps linked to their own speech production patterns (though we cannot test this hypothesis with the perceptual data). Because sonority expansion is not documented as being variable in this way in the literature, we might expect a more consistent impact of prominence across participants in Experiment 6. At the same time, it has been noted that lower vowels vary more acoustically as a function of sonority-expanding gestures as alluded to above (Babel, 2009; Beckman et al., 1992), in comparison to higher vowels where sonority expansion is "suppressed".[5] If general variability in a vowel category is larger for lower vowels we might expect that perceptual shifts as a function of prominence are more variable as well, based on exposure to more within-vowel-category variation, as compared to high vowels. Seeing how participants vary across experiments will allow us to explore these two predictions.

[4]As noted above, Cho (2005) also finds larger lip aperture for /i/ when prominent, suggesting that even though the tongue is fronted, some sonority expansion is occurring simultaneously, though clearly to a much smaller extent than for /ɑ/.

[5]More generally, it can be noted that lower vowels in American English tend to occupy larger F1/F2 spaces overall as compared to higher vowels (Clopper et al., 2005; Peterson & Barney, 1952).

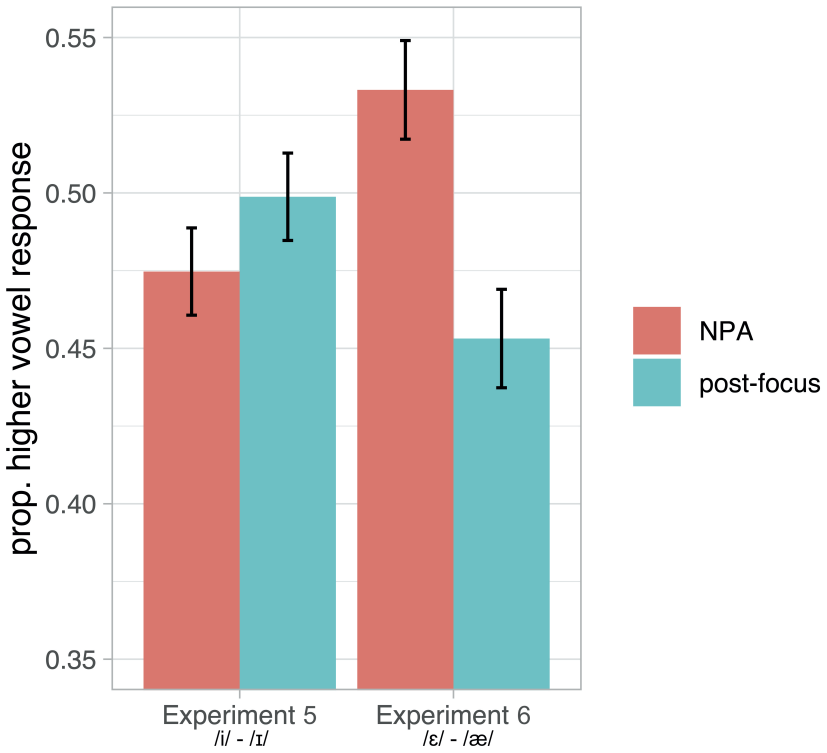

Figure 4.6: Comparison of the overall prominence effect in Experiments 5 and 6, with the proportion of responses for a higher vowel (/i/ and /ɛ/ respectively) plotted on the y axis, and categorization split by experiment and condition (at right). Error bars show 95% CI.

As a visual comparison of Experiments 5 and 6, Figure 4.6 shows listeners' overall responses in each experiment, pooled across continuum steps and split by prominence condition. One commonality across experiments is that the pair of vowels used in each varied in terms of height. In Experiment 6 (= Experiment 1) /ɛ/ is a higher vowel than /æ/, and in Experiment 5, /i/ is a higher vowel than /ɪ/. Across experiments, /ɛ/ and /i/ are vowels produced with more closed (or, less sonorous) articulations as compared to the other endpoint of the continuum. Accordingly, in Figure 4.6, /i/ and /ɛ/ are referred to collectively as "higher vowels", relative to /ɪ/ and /æ/ respectively.

As already noted, Figure 4.6 shows that the effect of prominence changes based on the vowel contrast in question. For /ɛ/ and /æ/ in Experiment 6, listeners adjust categorization such that a more sonorous vowel (where acoustically sonority refers to higher F1 and

lower F2) is expected in prominent contexts. The opposite is true in Experiment 5: a less sonorous vowel is expected in prominent contexts. Also apparent in Figure 4.6, the effect in Experiment 6 is larger in magnitude than the effect in Experiment 5.[6] This difference in the magnitude of each effect is consistent with the observation that sonority expansion entails larger articulatory and acoustic modulations, which might accordingly translate into larger perceptual adjustments for vowel contrasts which undergo sonority expansion.

Next, to quantify variability in each experiment, random slope estimates for the prominence manipulation in each model were inspected. The larger the estimated standard deviation is for the slope, the more variability across participants exists with respect to the prominence effect. In Experiment 5 the standard deviation of slope estimates is 0.13, while in Experiment 6 it is 1.34. These estimates therefore suggest that participants are substantially more variable in Experiment 6, as compared to Experiment 5.

To investigate further how participants varied across experiments, a by-participant estimate for the effect of prominence was obtained. This was calculated from best unbiased linear predictors (BLUPs) for each participant. In a mixed model, BLUPs represent how much an estimate differs for each participant (or item, where relevant) from the overall estimate for a fixed effect (Baayen, 2008; Blouin & Riopelle, 2005). By adding by-participant BLUPs to the fixed effect estimate, we obtain an estimate for the impact of a given effect on each participant, which notably factors in other effects in the model, as compared to e.g., by-participant differences between conditions (Politzer-Ahles & Piccinini, 2018). By-participant estimates, calculated in this way, accordingly let us visualize the consistency and distribution of an effect across participants. To this end, by-participant estimates were calculated for Experiments 1, 5 and 6. Figure 4.7 shows these estimates in addition to the fixed-effect estimate and 95% credible intervals from each model.

Note that in Figure 4.7, whether an estimate is positive or negative corresponds to whether the prominent NPA condition increased "higher vowel" responses, as it did in Experiments 1 and 6, or decreased them, as it did in Experiment 5. An individual participant's

[6]This is also apparent in comparing model estimates for the prominence effect in each experiment: in Experiment 6 $\beta = 0.49$, in Experiment 5 $\beta = -0.26$.

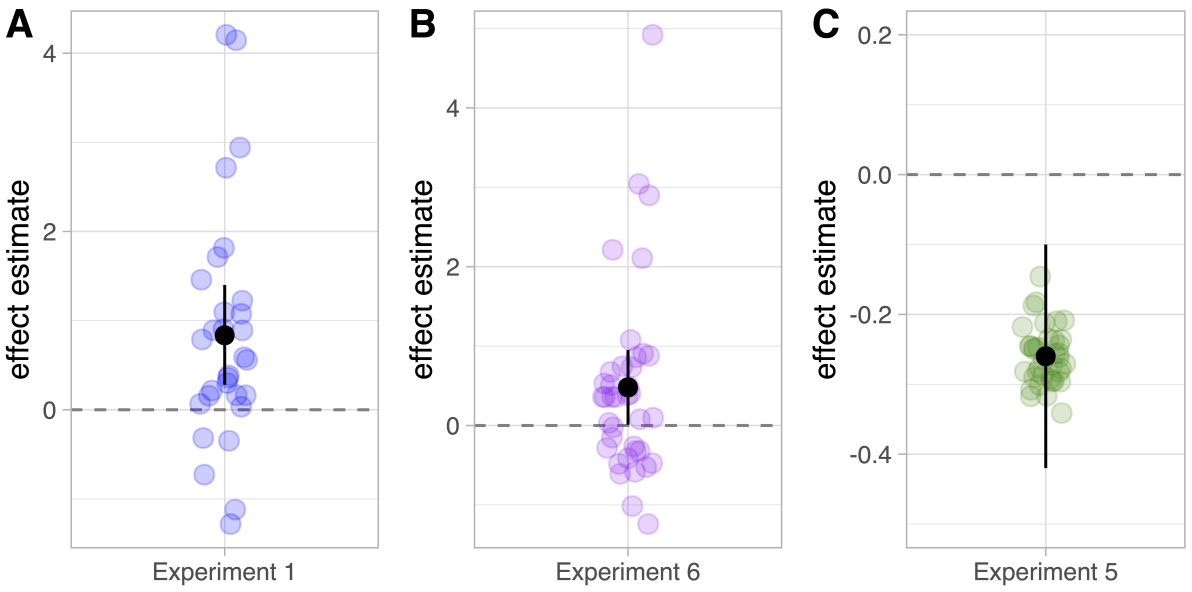

Figure 4.7: Comparison of by-participant prominence effects in Experiments 1 (panel A) 6 (panel B) and 5 (panel C). Colored points in each plot represent an effect estimate for each participant in log odds space (see text). The black circle and error bars show the model estimate for the prominence effect and 95% CI. Note that the y axes in all panels have different ranges.

estimate is the same: a positive estimate shows increased "higher vowel" responses, and a negative estimate shows decreased "higher vowel" responses.

First, a brief comparison of Experiment 1 to Experiment 6 can be considered (comparing panels A and B in Figure 4.7). As discussed above, the expected prominence effect was observed in both experiments. However, as can be seen in Figure 4.7, the effect estimate is clearly larger in Experiment 1 (β=0.83, 95%CI =[0.27,1.39]) as compared to Experiment 6 (β=0.49, 95%CI =[0.01,0.95]), though credible intervals for each estimate are fairly comparable in their range. This suggests that presenting stimuli remotely resulted in more participants showing a reversal of the main effect. As is apparent in Figure 4.7, more individual participants showed an effect that was the opposite directionality of the main effect in Experiment 6 (12 out of 38 participants, in fact), indicating that the effect is quite variable across participants. One speculative explanation for this is that participants showed

an increased reliance on durational cues in Experiment 6 as compared to Experiment 1. If in Experiment 6, overall audio quality was degraded and listening conditions were nosier, as compared to the high quality over-the-ear headphones and sound-attenuated room used in the lab, participants may have more heavily weighted duration as a cue to the /ɛ/-/æ/ contrast. This is plausible given that in noisy and degraded listening conditions, perception of spectral information is diminished while temporal information is relatively robust (Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995; Xu, Thompson, & Pfingst, 2005; Xu & Zheng, 2007). As noted in Section 2.2.2, reliance on duration could generate a reversal of the observed main effect due to the duration of preceding material and durational contrast effects (Diehl & Walsh, 1989; Newman & Sawusch, 1996). Further comparison between Experiments 1 and 6 is left aside here though from a methodological standpoint comparing other effect estimates across experiments in this way may be useful.

The central comparison of interest is that of Experiment 5 and 6, shown in panels B and C of Figure 4.7. As noted above, the prominence effect in Experiment 6 is clearly highly variable. In contrast, the prominence effect in Experiment 5, though small in magnitude, is *very* consistent (note the y axis on panel C of Figure 4.7). Unlike Experiments 1 and 6, no participant shows a reversal of the main effect and all participant estimates are clustered closely together. Thus inspection of by-participant estimates confirms what was noted in comparison of model slope estimates: participants are remarkably consistent with respect to the hyperarticulation effect in Experiment 5. In light of this finding we can therefore confirm that the results of Experiment 5 show a robust and consistent hyperarticulation effect, one that is fairly uniform across participants.

Why might we see such robust perceptual effects in the case of Experiment 5? If we assume that the variability surveyed in the speech production literature is really representative, one possibility is that this constitutes an asymmetry in how speakers produce prominence strengthening, and how listeners exploit its acoustic consequences, i.e. a production-perception "mismatch", in the vein of e.g., Harrington, Kleber, and Reubold (2013); Mitterer and Ernestus (2008). How might we explain this apparent asymmetry? We can consider these results in light of one of the functions of prominence strengthening, *paradigmatic con-*

trast enhancement. Given that, in Experiment 5, listeners are categorizing vowels as /i/ or /ɪ/, a prominent context might lead them to expect that these high front vowels will be maximally acoustically differentiated from other vowel categories in the language. In other words, when sonority expansion is less important, listeners might expect prominence to correlate with maximal acoustic differentiation or dispersion of vowels in the vowel space. This follows from the idea that sonority expansion is "suppressed" for certain vowels (Cho, 2005; de Jong, 1995), and when this is the case, acoustic dispersion is expected perceptually. In this sense we could define the contextual prominence effect in Experiment 5 as contrast enhancing, i.e. as leading to an expectation that these vowels will be maximally acoustically distinct in the vowel space when prominent (in relation to other vowels in the language). This is different from what we've seen in previous experiments, where enhancement of a vowel's sonority is expected when that vowel is prominent. Thus we have evidence for a clear difference in how prominence strengthening effects operate in perception, as a function of vowel features.

In summary, we can conclude the following from this comparison across experiments. Perceptual adjustments for hyperarticulation are small in magnitude, but highly consistent across participants. In comparison, perceptual adjustments for sonority expansion are relatively large in magnitude, but simultaneously quite variable. This speaks against the prediction that hyperarticulation effects may be more variable due to listeners' exposure to various patterns of prominence strengthening for high vowels. One possible explanation for the asymmetrical variability in each effect is that overall greater acoustic variability in the realization of sonority-expanding gestures in lower vowels (Babel, 2009; Beckman et al., 1992), leading to more variation in how listeners exploit their perceptual experience. If it is the case that "strengthened" non-high vowels show more acoustic variation than "strengthened" high vowels overall (even if high vowels like /i/ show various patterns of prominence strengthening in production), we might expect listeners to be more consistent with respect to the latter.[7] The present data cannot address this, though future work might benefit

[7]As another way of considering this asymmetry, it can be noted that mid vowels like /ɛ/ may lack a clear acoustic or articulatory target (i.e. specification of [-high] and [-low]). As such, the "sonority features" of the vowel (following de Jong, 1995) are prioritized, where for high vowels, a clearer acoustic target exists, and is prioritized in lieu of sonority.

from exploring questions such as these, by e.g., exposing participants to different degrees of stimulus variation prior to a categorization task.

Overall, the difference in both the magnitude and variability of the effects in Experiments 5 and 6 could be taken to show a link between perceptual prominence strengthening effects, and vowel-intrinsic properties, which are supposed here to arise from the ways in which different vowels can be strengthened by prominence. However, the "mismatch" between the various speech production patterns in the literature, and the perceptual result in Experiment 5 allows us to speculate that there is not a direct mapping between production and perception in this regard, though vowel-intrinsic features are clearly playing an important role.

4.6 General discussion

Findings in this chapter demonstrate that listeners are sensitive to hyperarticulation effects in speech perception, expecting more extreme F1 and F2 values (lower F1, higher F2), corresponding to a hyperarticulated high front vowel, in prominent contexts. The core finding in this chapter is accordingly that the effect of contextual (phrasal) prominence in vowel categorization is dependent on vowel-specific properties: prominence strengthening effects in perception are crucially dependent on how a specific vowel is strengthened, as opposed to being uniform across vowel categories. In this sense the present findings speak to the need for listeners to have access to what "counts" as a strengthened variant of a given vowel, which will vary based on vowel-intrinsic features. In other words, the perceptual processing observed throughout this dissertation must make reference in some way to vowel-specific features and the way they are strengthened prosodically.

We additionally observed that perceptual hyperarticulation effects are remarkably consistent across participants. This is in comparison to the variable sonority expansion effects, documented in Experiments 1 and 6. The consistency of the effect in Experiment 5 is somewhat surprising given the previously documented variability in the speech production literature. In the absence of speech production data from the participants in the current study, we can only speculate as to the extent the perceptual outcome aligns with their own

production repertoires and/or their perceptual experience more generally.

A more general idea is that perceptual prominence effects are language-specific and depend on the inventory and strengthening patterns of a given language. As mentioned above, Garellek and White (2015) found that, in Tongan, all vowels (including /i/), showed robust F1 raising when prominent. It would therefore be informative to test how Tongan listeners respond to contextual prominence manipulations of the sort employed here (in e.g.,categorizing a Tongan /i/ to /e/ continuum). On one hand, we might expect that Tongan listeners would show the opposite pattern reported here, such that more sonorous formant values for /i/ are expected in prominent contexts. On the other hand, it is possible that general acoustic dispersion will be expected (as seen in Experiment 5), such that more peripheral F1 and F2 values are required to perceive /i/ in prominent contexts. If the same pattern from Experiment 5 was observed for Tongan listeners, it would suggest a more general perceptual effect of prominence for vowels like /i/. In the Tongan case, this would be a clear mismatch with the observed acoustic effects of prominence. Seeing how closely perceptual prominence effects align with speech production patterns cross-linguistically, especially in comparing similar vowel categories, would help us better understand the extent to which these effects are linked to acoustic patterns in a given language. Testing the extent to which these effects are further influenced by the vowel inventory of language, in comparing, e.g., a less crowded vowel space in Tongan (where dispersion may be less of a priority) to the English case tested here, would further be informative.

CHAPTER 5

Discussion and conclusion

5.1 Overview of findings

The experiments contained in this dissertation revolved around the following questions which were outlined in Chapter 1. Below they are reiterated and answered.

5.1.1 Does prosodic prominence mediate perception of vowel contrasts?

We can answer this question with a clear "yes". Throughout the experiments contained in this dissertation we have seen evidence that contextual prominence shifts how listeners map formant cues to a vowel category. The observed perceptual adjustments generally fit with the way in which formant structure is modulated by prominence. This finding extended some of the past work on prosodic context effects to test how spectral cues (instead of temporal cues) are perceived by listeners. It is also the first evidence to my knowledge to show prosodic effects on the perception of vowel quality contrasts, where previous comparable studies have tested vowel length and voicing contrasts (Kim & Cho, 2013; Steffman, 2019b; Steffman & Katsuda, 2020). Whereas these previous studies focused on the boundary-marking function of prosody, the present experiments show contextual prominence also plays a role in segmental perception. The question of whether the observed shifts in vowel categorization that we have seen can be explained by durational/spectral contrast is an additional point of interest, as discussed in Section 1.4.3 (see also Mitterer et al., 2016; Steffman, 2019a). The experiments throughout this dissertation showed that this is not the case for the effects observed here, because durational contrast predicted the opposite of the observed shift in categorization in phrasal contexts, and spectral contrast predicted the opposite of the glottalization

effects seen in Chapter 3.

The timecourse data for phrasal prominence, discussed below, also suggests that these effects are not the result of domain-general contrast. Instead, they are hypothesized to arise from listeners' reference to prosodic structural context, or a parsed out prosodic structure following Cho et al. (2007).

5.1.2 How is prominence integrated with segmental cues?

The processing results in this dissertation highlight, fundamentally, a multi-stage role for prominence in processing. Experiment 2 showed that phrasal prominence was most influential later in processing, with the effect reaching its maximum at the end of the analysis window. The prominence effect was overall asynchronous with the influence of formant cues, which occurred early in processing, as assessed by when smooths in the model diverged. Both the overall delayed effect and the asynchrony with listeners' use of formant cues are consistent with the two-stage prosodic analysis model set forth on the basis of prosodic boundaries (Cho et al., 2007; Kim, Mitterer, & Cho, 2018; Mitterer et al., 2019). This supports the idea that phrasal prominence is processed by listeners as an abstract (phonological) prosodic structure that enters into the later stages of spoken word recognition. However, at the same time, it was apparent that phrasal prominence exerted a subtle early influence. Listeners shifted their perception of formant cues based on prominence condition such that, at the earliest point in time they showed a preference for one target or the other, the same F1/F2 values on the continuum were perceived as more or less like /ɛ/ or /æ/ as a function of prominence. In other words, listeners' use of formant cues was impacted by prominence in the earliest stages of processing. This runs counter to the idea that prominence is *only* a later stage influence.

The two accounts that were proposed to explain prosodic and segmental integration therefore both appear to be right, to a certain extent. Immediate effects of prominence on formant processing are taken to originate from the acoustic/phonetic prominence of the target sound, while the fact that the effect is strongest later in time, and attains its maximum late in pro-

cessing, is in accordance with the idea that prosodic context mediates selection of activated lexical candidates, following Cho et al. (2007), Kim, Mitterer, and Cho (2018) and Mitterer et al. (2019), and in line with the idea that structural prosody may take time to compute. The key takeaway from this finding in Experiment 2 is thus that phonetic prominence can exert early influences in processing, but nevertheless overall delayed prominence effects are consistent with prosodic analysis. Prominence information therefore crucially enters into processing at multiple stages, a point discussed further below.

5.1.3 Does segmental context (glottalization) cue prominence?

The experiments in Chapter 3 showed that glottalization shifts listeners' perception of vowel contrasts in comparable fashion to phrasal prominence. When a word is preceded by [ʔ], listeners expect it to contain a vowel with more sonorous formant values, the same effect as was seen based on the phrasal prominence manipulation in Chapter 2. As stated above, this effect was shown to be independent of spectral contrast, suggesting again that prominence-lending context is exerting an independent influence. The experiments in Chapter 3 thus show that prominence effects can be elicited in the absence of a manipulated phrasal context. More generally, they suggest that prominence-lending cues in perception include those that reliably co-occur with, or bundle together to signal, phrasal (phonological) prominence. In this view, glottalization is not a phonological/structural prominence category, but serves this perceptual prominence-marking function to listeners because of the way it patterns with prosodic structure (or, the way it encodes prominence; Garellek, 2013, 2014). The findings in Chapter 3 accordingly offer support for the claim that voice quality, particularly glottalization, plays an important role as a prominence marker, by showing that listeners treat it as such in segmental perception.

As discussed in Chapter 3, these results more generally predict that other localized prominence strengthening effects, such as prominence strengthening on consonants, should show a similar impact on vowel perception. We can predict that lengthened VOT, nasal duration, or fricative duration (Cho et al., 2017; Cole et al., 2007; Silbert & de Jong, 2008) would

lead to the expectation of prominence strengthening in the formants of a following vowel. We can also predict such effects might go the other way. For example, formant structure in a following vowel could impact listeners' perception of a VOT continuum, as a function of how lengthened VOT and sonorous formant structure co-vary to mark prominence. Seeing if effects such as these occurred in words in isolation would be a further test of their scope, i.e. if segmental strengthening effects have an impact in perception even when more global prosodic context is missing. Because Experiments 3 and 4 manipulated *only* the presence or absence of a glottal stop, we may already have evidence that this is the case. This lines up with Steffman and Jun (2019), who found prominence-driven effects of pitch height in vowel duration perception in isolated words. If future results further support this idea, we would have evidence that prosodic structural co-variation in cues shapes perception in a structure-independent way. In other words, acoustic cues which pattern together as a function of abstract prosodic structure, or more generally cue phonetic prominence, can impact perception even when extended prominence-lending phrasal structure is absent in the signal (i.e. in isolated words). Evidence along these lines would show just how far-reaching effects of prosody in speech perception can be, in line with its conception as a central organizing force in how speakers plan and execute speech (Beckman, 1996; Cho, 2016; Keating & Shattuck-Hufnagel, 2002). These ideas are further discussed in Section 5.2.

5.1.4 Does prominence processing vary based on prominence-lending context?

This question was addressed by comparing eye movement data in Experiments 2 and 4, which tested the effects of phrasal prominence and glottalization respectively. The clear answer to this question is "yes". As discussed above, the effect of phrasal prominence could be characterized as multi-stage, and clearly different from the effect of formant cues, both in terms of the time in which it impacts processing robustly, and the growth of the effect over time. In contrast, the timecourse of the glottal stop effect was essentially synchronous with the uptake of formant cues, with both occurring early in processing. If a glottal stop is taken to constitute a local (or, segmental) cue to prominence, we can conclude that local prominence-lending context is integrated immediately with segmental cues, lining up with

the general timecourse we'd expect for compensation for e.g., preceding speech rate or spectral context (Maslowski et al., 2020; Reinisch & Sjerps, 2013; Toscano & McMurray, 2015). This timecourse is taken to reflect listeners' sensitivity to the co-variation of prominence signaling acoustic properties (including variation within phonological prominence categories, as in Experiments 3 and 4). That is, based on the way they pattern together to cue prominence, [ʔ] generates an immediate expectation for how a vowel should be realized, and immediate perceptual compensation, following e.g., McMurray and Jongman (2011); Toscano and McMurray (2015). These "phonetic" prominence effects are therefore taken to impact processing in a way that does not involve explicit prosodic analysis. This contrasts with the multi-stage effect of phrasal prosodic structure outlined above, which showed its strongest influence later in processing. What both glottalization and phrasal prominence share is an early-stage fine-tuning of formant perception, via phonetic prominence, but where they differ crucially is their overall timing with respect to the influence of formants, and in how each effect grows over time.

5.1.5 Do perceptual prominence effects vary based on vowel-intrinsic features?

Experiments 5 and 6 addressed this question by comparing how different vowel contrasts were impacted perceptually by phrasal prominence. The expectation for how a vowel should be realized in a prominent context was observed to vary based on the vowel in question (specifically, based on vowel height). More extreme (hyperarticulated) F1 and F2 values were expected for high vowels /i/ and /ɪ/. The opposite modulation in F1 and F2 was expected for /ɛ/ and /æ/, in line with sonority expansion. Interestingly, despite various patterns attested for prominence strengthening in /i/ in the speech production literature, listeners showed a consistent and robust expectation of (acoustic) hyperarticulation. At the most basic level, this shows that perceptual prominence effects generate an expectation of various strengthening patterns, though as discussed in Chapter 4, the extent to which these differences are directly linked to speakers' own production repertoires is unclear.

We can contrast the observed outcome with a hypothetical one in which, regardless of

vowel category, the same modulations in F1 and F2 were expected by listeners in prominent contexts. This outcome would suggest a more general perceptual mechanism that applied the same compensatory adjustments when any vowel was cued as prominent. Instead, we have evidence that the perceptual mechanisms at play make reference to vowel-specific features in generating expectations for how prominence should shape vowel realization. One possibility for these differences, forwarded in Chapter 4, is that for vowels which do not reliably undergo sonority expansion, peripherilization in the vowel space is expected under prominence. For vowels that *do* undergo sonority expansion in a systematic way, sonority expansion effects are dominant. Future work will benefit from further testing these ideas, particularly in testing how the vowel inventory of a given language might mediate these effects (where perceptual dispersion would be most useful in crowded vowel spaces), as discussed in Section 4.6.

5.2 Towards a model of prominence and segmental processing

The present results are only the first step towards explicating a complete and predictive theory of prominence effects in segmental processing (and processing more generally). However, they do provide some important constraints on the processes responsible, and desiderata for a model. Some core insights are outlined below, after which the architecture of a model is proposed.

5.2.1 Prominence processing as pre-lexical

As described in Chapter 1, the current theory of prosodic and segmental processing holds that listeners extract parallel prosodic and segmental representations from speech as it unfolds, but that the stage at which prosodic and segmental information contribute to word recognition is different (Cho et al., 2007; Kim, Mitterer, & Cho, 2018; Mitterer et al., 2019). The segmental analysis activates lexical hypotheses, while the prosodic analysis is integrated via lexical competition, that is, after contact with the lexicon is made. The model thus delimits prosodic influences in decisions about segments as post-lexical. As discussed in Chapter 1, there is clear empirical support for this idea, as shown by online data from

phrasing-modulated phonological inferencing in Korean (Kim, Mitterer, & Cho, 2018) and glottalization as boundary marking in Maltese (Mitterer et al., 2019).

The present dissertation has shown that, in various ways, contextual prominence influences are *not strictly post-lexical*. As such, one central contribution of this dissertation is that prominence information can be integrated early in processing by listeners. Though, as will be discussed in Section 5.2.2, the present results are seen as being compatible with the prosodic analysis model, they add more nuance to our understanding of the influence of prosody in word recognition in showing that its influence needn't be restricted to later stage (post-lexical) processing.

How should prosodic prominence enter into the earlier stages of speech recognition? The patterns we see in both Experiments 2 and 4 could be seen as reflecting *immediate compensation* for contextual prominence. The listener's construal of this sort of phonetic prominence has also been shown to incorporate various cues. Perception of prominence of a given linguistic unit will be closely linked to its salience in relation to its context (Mo, 2011), dependent on any acoustic properties that help set it apart from neighboring material, such as increases in f0, intensity and duration (all shown to predict prominence perception in different languages in e.g., Cole et al., 2019). At the same time, phonetic prominence is likely conveyed by language-specific patterns of segmental co-variation (stopping of fricatives, changes in vowel formant frequencies, and so on). The present findings show clearly that a model which desires to explain how prominence impacts processing must be "phonetically informed", that is, it must be based on a solid understanding of how prominence is encoded phonetically in a given language.

The observed immediate impact of prominence on the perception of formant cues seen here is compatible with a mechanism such as perceptual re-coding of cue values in a model like C-CuRE (McMurray et al., 2011; McMurray & Jongman, 2011), which also has the desirable property of representing a cue value in phonetic detail after compensation (see discussion in Section 1.4.1). As such, the perceptual mechanisms at play in the early stages of prominence processing are likely accounted for by an existing understanding of how listeners integrate cues with context. The contribution of the present experiments in this regard is in expanding

the notion of "context" to include phonetic prominence, which has been shown to clearly impact how segmental cues are perceived. Accordingly, a full understanding of how context influences segmental processing must include an understanding of prosodic prominence (in a given language), the way it is encoded phonetically, and the extent to which prominence-lending phonetic cues impact perception of one another.

More generally, these findings highlight the need to consider phonetic detail in modeling prominence effects in processing. The full relevance of phonetic detail in higher-level processing related to prominence remains to be seen (for example in resolving contrastive discourse referents and informational structural constraints, as in a task such as e.g., Ito & Speer, 2008). However, the present results show that we should certainly consider phonetic detail as relevant throughout processing. In agreement with various previous studies, these results show that modeling prosodic prominence in a strictly abstract/symbolic fashion will miss a part of the picture (see also Baumann & Cangemi, 2020; Cole & Shattuck-Hufnagel, 2016; Grice et al., 2017; Roessig et al., 2019).

5.2.2 Prominence as phonological structure

Based on the discussion in Section 5.2.1, one might be tempted to discard the prosodic analysis model for prominence effects, under the assumption that prominence perception can be reduced to phonetic cues which do not feed forward to a more abstract prosodic parse. Though there are independent reasons to be skeptical of this possibility, the present results speak against it as well.

A "purely phonetic" account of prominence perception that does not make recourse to more abstract prosodic organization would be unable to explain the notable differences in processing observed between Experiments 2 and 4. Though in both cases the prominence manipulation exerted an immediate influence on the perception of formants, the timecourse assessment showed that a reliable difference between prominence conditions occurred later in time when prominence was cued phrasally. This delay was observed in relation to the uptake of formant cues (comparably rapid in both experiments), and in comparing the prominence

effects across experiments, as shown in Figure 3.8. In a model where prominence perception is strictly *pre-lexical*, we should expect both of these effects to show the same timecourse and trajectory (though of course they may differ in their magnitude). As described in Section 3.5, this difference across experiments is supposed to arise from the incorporation of phonetic prominence with phonological prominence in Experiment 2, where phrasal context varied to convey differences in (phonological) accentuation.

This interpretation of the results accords with the prosodic analysis model in positing that (1) phonological prominence information is extracted in parallel to segmental information (like prosodic boundaries, as discussed in Cho et al., 2007; Mitterer et al., 2019) and that (2) this prominence information enters into processing at a later stage. In this sense the present results also fit with the existing theory of prosodic and segmental processing. However they extend these ideas to capture the fact that prominence information (not just prosodic boundaries) should enter into post-lexical processing, and that this prosodic parse of the signal is preceded by listeners' integration of phonetic prominence cues.

These results more generally fit with the idea that prosodic structure needs to be parsed "in its own right" (Beckman, 1996; Cho, 2016), in showing that phonological prominence is brought to bear on word recognition. Without a conception of phonological prominence as a component of linguistic structure, we would be unable to explain the effect of prominence in Experiment 2, and particularly its delayed timecourse. The precise mechanism of post-lexical integration (that is, the relationship and processing interactions between prosodic structural information and lexical candidates) remains to be explored. Cho et al. (2007) conceive of boundary information from prosodic analysis as influencing lexical processing via alignment of prosodic boundaries and word boundaries (see also Salverda, Dahan, & McQueen, 2003, cf. Christophe et al., 2004). This sort of prosodic boundary information would be only partially relevant in relating prominence encoded in parsed prosody to lexical candidates (e.g., in determining the status of accents as nuclear in relation to phrasing). We do know that whatever mechanism of integration is operative, it must be "phonetically informed" in having access to what variants of word forms count as strengthened, as shown in Chapter 4. These questions are outside the scope of the present dissertation, though one possible

model for considering these prominence effects in prosodic analysis is one in which multiple pronunciation variants for a given word form are stored in the mental lexicon, a notion that has empirical support from studies involving reduced speech and other variation in how words are realized (Arndt-Lappe & Ernestus, 2020; Brand & Ernestus, 2018; Ernestus, 2014, cf. Schweitzer et al., 2015).[1] The process of integrating prominence information in lexical competition would thus involve mapping prosodic structural information (i.e. is this word accented? is this word in nuclear position in a phrase?) to pronunciation variants in the lexicon that varied in their prominence (i.e. an accented word versus an unaccented word). In that sense, the prosodic parse of the signal would factor into lexical competition by matching (phonological) prominence information with the appropriate lexical candidate. This remains to be explored in future research.

5.2.3 Prominence as facilitation

One issue that has been touched on only very tangentially in the present experiments is the idea that prosodic prominence should facilitate speech recognition. Following the idea that prominence marking correlates with informationally rich or important linguistic material (Baumann & Cangemi, 2020; Ladd, 2008), it makes sense that prosodic prominence should facilitate word recognition and language comprehension more generally, as has already been shown by numerous previous findings (Baumann & Schumacher, 2020; Cole & Jakimik, 1980; Cutler et al., 1997; Cutler & Foss, 1977; Shields, McHugh, & Martin, 1974). Perhaps most relevant to the present experiments, particularly those experiments which manipulated phrasal prominence, L+H* accentuation has been shown to facilitate phoneme monitoring (Cutler, 1976; Rysling, Bishop, Clifton, & Yacovone, 2020): listeners are quicker to detect a phoneme in speech when the word with that phoneme is accented as compared to unaccented. Rysling et al. (2020) further recently showed that cohesion with the syllable preceding an accented target word is necessary for this effect to obtain, that is, accentual prominence perception clearly incorporates contextual information (e.g., falling pitch approaching the

[1] As discussed in Ernestus (2014), information about pronunciation variants could coexist with more abstract representations in so-called hybrid models.

low target in L+H*).

Throughout the experiments reported here we saw various ways in which prominence facilitates speech perception. In categorization responses, a prominent context led to stronger anchoring in categorization functions, suggesting better discrimination of F1/F2 differences. In online processing, prominence led to more rapid preferences for a visual target, as shown by the surface plots. These findings are perhaps not surprising, though they indicate that prominence as facilitation extends to both how acoustic cues are perceived, and how quickly they contribute to word recognition. A model of segmental processing and word recognition that seeks to capture this aspect of prominence strengthening in perception will need to encode not only the fact that prominence shifts a perceived cue value, but also modulates the speed at which listeners use a cue. This reinforces the idea that, as shown throughout this dissertation, listeners are not only determining which segmental contrasts they are hearing as speech unfolds, they are also influenced by how prominent these units are. In this sense the present findings offer a basic extension of the notion of prominence as facilitation to (1) the perception/categorization of acoustic cues, and (2) the processing of acoustic detail, including at the pre-lexical level.

5.2.4 Towards a model: The MAPP proposal

To tie these ideas together, we can consider the schematic architecture of a model that can account for these findings, shown in Figure 5.1. In reference to the main finding of this dissertation, we'll call this proposal MAPP, for Multi-stage Assessment of Prominence in Processing. In this section we'll walk through the schema in Figure 5.1, with a focus on how it accounts for the findings in this dissertation, and we will subsequently consider how certain aspects of the proposal could be further tested in future research.

The general architecture of the model is a layered network of various units, where information is passed between different levels of representation, as is fairly standard in models of perception and word recognition (Luce & Pisoni, 1998; Marslen-Wilson & Tyler, 1980; McClelland & Elman, 1986; Norris, 1999; Norris & McQueen, 2008). The architecture in-

corporates network activation principles of lexical activation and competition as common in models of spoken word recognition, though as will be discussed, the model is agnostic with respect to the precise implementation of lexical competition and related processing interactions. The model as described here is conceived of as being structured around modular and feed-forward flow of information, though it is easy to imagine a variant with interactive feedback in certain parts (in the vein of, e.g., TRACE; McClelland & Elman, 1986), a point that is tangential to the main claims here which will be discussed briefly below. The schema in Figure 5.1 is decomposed into several "layers", labeled A-D. These will be used to refer to certain parts of the model in the following discussion.

In outlining the architecture of the proposal, we'll take the schematic example shown in Figure 5.1. The process begins at a given time slice where acoustic information in the form of cues (indexed C_1 through C_5) is extracted from an acoustic event, as shown in layer A. In the example given in the figure, these cues are coming from the boxed vowel in the waveform shown at the bottom. Note that this vowel is relatively prominent in terms of amplitude and pitch, a point that will be relevant in illustrating some features of the proposal.

At layer B, note that cues are fed forward in two directions, towards a prosodic analysis (at right), and a segmental analysis (at left). The usefulness of a cue for the segmental and prosodic analysis will vary by language and the mapping will not be one to one: certain cues will certainly be useful for specifying both (e.g., duration and pitch, see discussion in Cho et al., 2007). This is represented in the schema by showing some cues feed towards the segmental analysis, some feed towards the prosodic analysis, and some feed towards both. At layer B, listeners construct a continuous/phonetic representation of prominence. We could think of this as a sort of sliding scale, as shown in the figure. This continuous representation of prominence is constructed from a combination of prominence-lending cues. Of course, these cues will be perceived relative to context (e.g., pitch will be perceived as high in relation to preceding pitch), and will further incorporate information like glottalization, segmental strengthening patterns, etc. This phonetic prominence information is integrated with cues which are feeding towards segmental analysis (shown by the dashed arrow which points leftwards), something we could model as cue re-coding as in C-CuRE (McMurray et

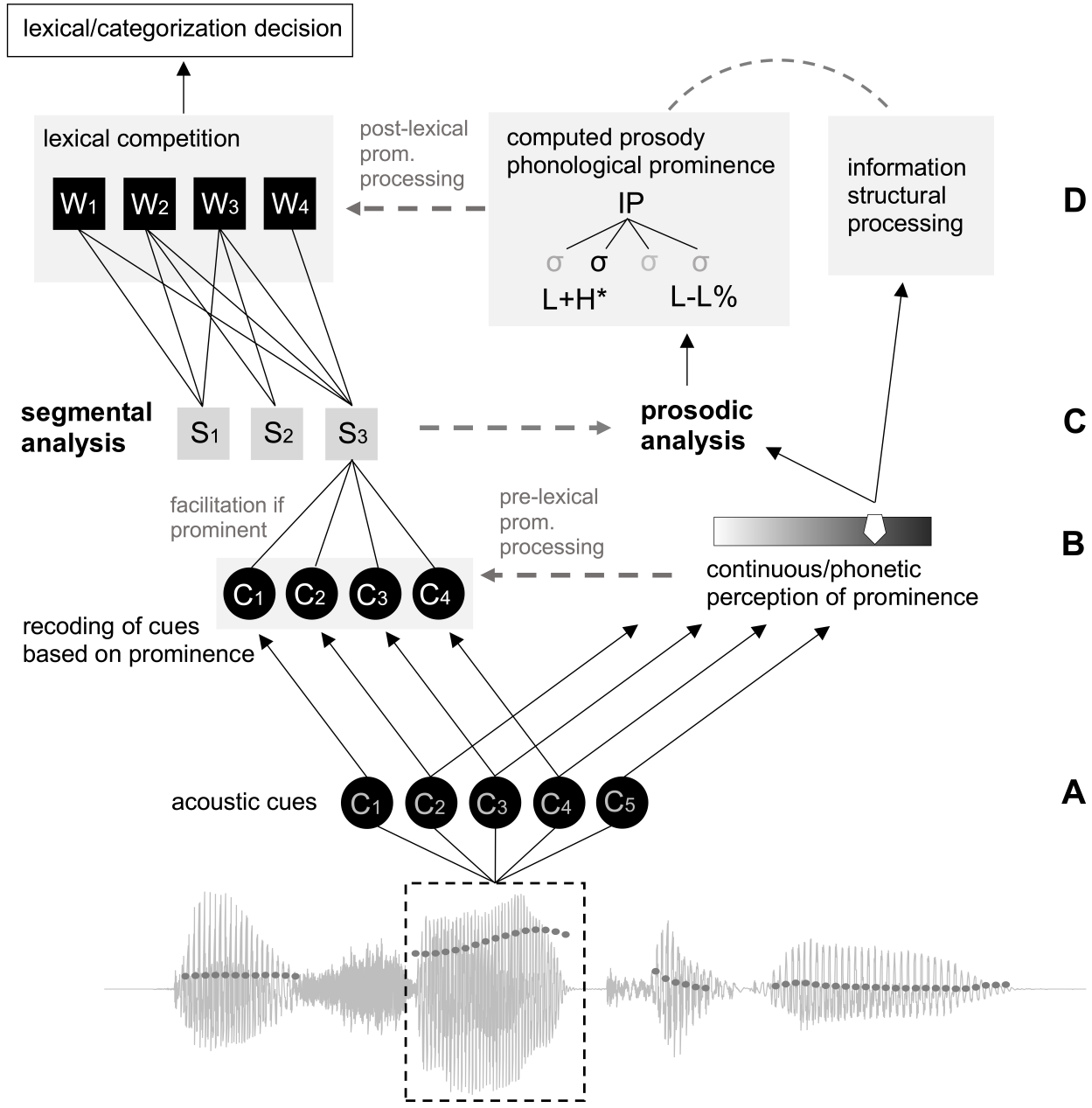

Figure 5.1: A schematic of the MAPP (Multistage Assessment of Prominence in Processing) proposal, showing the flow of information from signal to lexical access. Note that nodes with Cs, Ss and Ws, stand for "cue", "segment" and "word" respectively. See text for details.

al., 2011; McMurray & Jongman, 2011). This would entail retention of phonetic detail, but an altered percept based on prominence-generated expectations (note the color of the text of C_1-C_4 is changed from gray to white to represent this).[2] It is this sort of process which is assumed to operate in the early stages of processing in Experiment 2, where phonetic prominence immediately impacted formant perception. This is also assumed to explain the pattern seen in Experiment 4, where glottalization similarly impacted listeners' perception of formants immediately.[3]

In layer C, segmental cues and phonetic prominence feed into segmental and prosodic analyses. In this example, let us assume that all cues map straightforwardly onto a segmental category S_3, which is the third in a string, preceded by already determined categories S_1 and S_2. Note the dashed arrow pointing rightwards. This represents that segmental analysis might inform the prosodic analysis by providing information about e.g., intrinsic vowel length and voicing features which alter pitch/duration, and so on (Cho et al., 2007). Note too that prominence will facilitate the use of cues, speeding up, and increasing certainty about, the mapping from cue to segmental category (this part of the model would be consistent with aforementioned phoneme monitoring studies, Cutler, 1976; Rysling et al., 2020). A possible test for this facilitation component of the model could come from experiments which test how prominence leads to increased performance in discrimination tasks, both in terms of speed and accuracy.

---

[2]We can also speculate that this sort of pre-lexical prominence effect will capture segment-specific expectations of prominence strengthening. As cues combine and are processed to give listeners information about segmental features, expectations for the realization of a given cue could be modulated in a different fashion by prominence information, e.g., expectations for F1 in a prominent /i/ will be different than expectations for F1 in a prominent /ɛ/, as seen in Chapter 4. However, the extent to which segment-specific information is relevant in pre-lexical processing cannot be fully determined on the basis of the present results (e.g., in the absence of an experiment testing glottalization effects on high vowels).

[3]One other prediction of this part of the model is that, because a given cue can impact perceived prominence and also feed towards a segmental analysis, perceived prominence (based on a cue) could impact that cue's contribution to segmental processing. Interestingly, Steffman and Jun (2019) find support for this idea in their finding that vowel duration on a continuum that cued coda stop voicing *also* mediated the effect of pitch (as a prominence cue). Only when vowel duration was shorter did listeners interpret low pitch as cuing a lack of prominence, which led to compensatory perception of vowel duration as a cue to voicing. When vowel duration on the continuum was longer, the target was perceived as prominent on that basis alone, such that low pitch no longer cued a lack of prominence to listeners (instead pitch showed the opposite effect based on psychoacoustic integration with duration). This shows clearly that duration simultaneously contributed to prominence perception and cued a segmental category. See Steffman and Jun (2019) for details.

Additionally, note that another vertical arrow goes upwards from the continuous representation of prominence in layer B, bypassing layer C. This represents that phonetic prominence-lending cues needn't necessarily feed into abstract prosodic analysis, as we saw in Experiment 4 (see also Steffman & Jun, 2019). This also presents a possible route for phonetic information to modulate higher-level processing where prominence is relevant, consistent with e.g., Grice et al. (2017) who found that listeners relied on phonetic parameters to make decisions about a speaker's intended intonational meaning, conveying focus structure (cf. Fowler & Housum, 1987). Grice et al. (2017) found that phonetic information was relevant even within a pitch accent category (see also Cangemi & Grice, 2016; Roessig et al., 2019 for similar ideas), in line with this part of the model where phonetic prominence information can inform higher-level prominence processing.

Finally, in layer D, a parsed prosodic structure results from prosodic analysis, and is integrated with segmental information via lexical competition (that is, post-lexically) as shown by the gray dashed arrow pointing leftwards. The structure parsed from prosodic analysis will involve incorporation of phonetic prominence with many other pieces of information. In the example in the figure, a complete prosodic analysis would include integration of prominence in the boxed target sound in the waveform with structural/phrasing information, encoding the fact that it is the only pitch accented word in the phrase (and therefore is in nuclear position), that it bears a prominent L+H* pitch accent, and so on.

For obvious reasons, a complete prosodic analysis will necessitate integrating both preceding and following context, though it is proposed here that listeners will begin constructing prosodic structure as soon as they can, with processing that is incremental (e.g., Baumann & Schumacher, 2020; Cho et al., 2007; Schafer, 1997; Speer et al., 2003). As speech unfolds, listeners will then refine and modify a built structure as more information becomes available to them. For example, at the end of the target segment in this example, listeners will have access to pitch accent information, but will not yet know this is the nuclear pitch accent. A test for this sort of structure-dependent processing could come from cases where structural/phonological prominence information is uncertain early on, and only becomes clear to

listeners later in time.[4]

At the left of the figure in layer D, the relationship between segments and the lexicon is shown by links from decided-upon segmental categories to word forms in the lexicon (indexed with W_1 through W_4). In the schematic example, a string of three segments contributes variably to the four word forms under consideration, i.e. showing bottom-up support for a word (which we could model as activation). W_1 is supported by S_1 and S_3, W_2 is supported by all three segments, and so on. It is at this stage of lexical competition that various other pieces of information (e.g., neighborhood density) will be integrated in processing, and other dynamics of lexical competition such as inhibition (e.g., Norris et al., 2000; Vitevitch & Luce, 1998) will be relevant. As noted above, the MAPP proposal is agnostic with respect to the specific architecture of lexical competition. For example, if we wanted to allocate a module specific to phonemic decision making, as modeled with decision nodes in Merge (Norris, 1999; Norris et al., 2000), we can imagine that this would interact with segmental and lexical items as specified in that model (with prosodic analysis feeding into decision nodes via the lexicon). The architecture of the present proposal is conceived of as being feed-forward, in the sense that information does not flow backwards from higher to lower layers,[5] though the main ideas proposed here are not incompatible with a model that includes feedback, e.g., a feedback loop from lexical representations to segments, as in TRACE (McClelland & Elman, 1986).

Also note that in layer D, abstract prosodic structure and information structure processing are assumed to interact bidirectionally, that is, informational structural context might contribute to a parsed prosodic structure (e.g., Bishop, 2012, discussed below), while simultaneously, abstract prosodic organization will convey information about focus marking, given-ness, informativity, and so on (Calhoun, 2007; Cole, Mo, & Hasegawa-Johnson, 2010;

[4]For example, imagine a case where the number of unaccented syllables following an accent (before the end of a phrase) varies. When more unaccented syllables follow, listeners will have to wait longer to know whether an accented syllable has the nuclear accent (though this could be disambiguated by certain pitch patterns, i.e. an immediate unaccented low pitch target L- signaling no more accents will follow before L%). Seeing if the influence of prosodic analysis is delayed as a function of following syllables would be a useful test of the idea that structural information which takes longer to resolve leads to a greater delay in processing.

[5]This is a controversial issue that is not directly relevant to the present proposal: see e.g., Norris et al. (2000, 2018) for arguments against feedback, and commentary in Norris et al. (2000) as well as Magnuson et al. (2018) for arguments in favor of feedback.

Watson, Arnold, & Tanenhaus, 2008).

Finally, as the result of both of these paths in processing, listeners will make a lexical/categorization decision (note "categorization" is included here to be agnostic as to how this stage might operate in a decision making module as in a model like Merge). At the same time, listeners will use a parsed-out prosodic structure for other domains of processing, and may use phonetic/continuous prominence information as well, as discussed above.

It is worth outlining too what the schematic in Figure 5.1 does *not* show. A syntactic module of processing is not shown here; nevertheless this architecture could be co-opted to include syntactic processing, which would interact with prosodic processing in various ways, following e.g., Schafer (1997); Speer et al. (1996, 2003). Syntactic structure could also be modeled as factoring into lexical competition, following Swinney (1979); Tanenhaus, Leiman, and Seidenberg (1979), exerting a later-stage influence similar to prosodic analysis. Also not shown in the schematic is subsequent processing after lexical selection, i.e. lexical integration in the sense of Friederici, Steinhauer, and Frisch (1999). This will include integration of lexical information with syntactic and semantic information (Friederici et al., 1999; Swinney, 1979; Tanenhaus et al., 1979). Lexical selection and integration will also influence information structural processing, e.g., knowing if a word is new to a discourse (as compared to given) requires integrating a selected lexical item (see e.g., Baumann & Schumacher, 2020).

To sum up, in outlining MAPP we have seen how both the results of the present experiments and other work reviewed above (Kim, Mitterer, & Cho, 2018; Mitterer et al., 2019; Steffman, 2019b; Steffman & Jun, 2019) fit with its architecture. As is evident in the name, the scope of MAPP is rather limited; the main goal is to provide a model in which prominence information is processed in multiple stages, with a plausible structure in line with what we know about spoken word recognition and prosodic analysis.

MAPP makes various predictions. In the following section, some further directions are outlined, which will include suggestions of how the proposal can be tested in future work.

5.3 Further directions

As always, the results discussed above raise various questions that will benefit from future research. Below several broad areas of further study are outlined.

5.3.1 Further tests of pre-lexical prominence effects

Tests of the pre-lexical integration component of the model could take various forms. Further online measures showing an early timecourse for prominence integration (with e.g., eyetracking) would support the idea that prominence is being processed pre-lexically. As another test, methods which attempt to disrupt processing by e.g., manipulating attentional resources or cognitive load during a task (following e.g., Bosker et al., 2017) could tap into whether prominence processing is early or late. If we see that modulations in attention and cognitive load do not strongly diminish prominence effects in segmental perception, we would have further evidence for an early stage influence, i.e. one that is not modulated by other cognitive processes (where by contrast, we might predict prosodic analysis is disrupted by these manipulations).

Another test could come from continuous manipulations in a prominence-lending cue, with the goal of seeing if they lead to continuous shifts in segmental categorization (reflecting perception of phonetic prominence). Seeing the extent to which continuous changes in e.g., pitch, engender continuous or categorical shifts in segmental categorization could be used to probe how categorical listeners' representation of prominence is, given the now established link between prominence and segmental perception. Continuous shifts in categorization as a function of continuous changes in a prominence-lending cue would support the pre-lexical module of prominence processing in MAPP. The many ways in which prominence is encoded in segmental detail further offer a wide array of possible tests for how prominence-lending cues come together in segmental processing. Establishing the (ir)relevance of properties of interest could accordingly build a theory of what sort of acoustic information (in addition to f0, duration, and intensity) lends prominence, and a possible hierarchy of cues in this regard.

As a related extension, continuous prominence manipulations could be used to test repre-

sentational theories of intonation/prosody. Put in terms of MAPP, this would entail testing the extent to which phonetic prominence cues feed into abstract prosodic analysis. Imagine for example an f0 continuum ranging from an American English H* to L+H* pitch accent, where the latter is generally more prominent (Bishop et al., 2020). The existence and function of bi-tonal pitch accents and particularly L+H* has been a persistent controversy in theories of intonational representation (e.g., Dilley & Heffner, 2013; Dilley, Ladd, & Schepman, 2005; Ladd, 2008). Moreover, continuous within-category variation in tonal alignment and scaling are clearly important for intonational structure, blurring the lines between categorical/continuous intonational features (Grice et al., 2017). A segmental categorization task could thus be used as a test for listeners' perception of prominence itself, and in particular, different prosodic/intonational categories which are assumed to vary in their prominence. For example, testing if segmental perception tracked continuously, or shifted categorically, with changing pitch would be a test for the existence of discrete pitch accent categories in this case. More generally, testing how segmental perception tracked with continuous changes in other prominence-lending cues would be informative. For example, if continuous prominence-lending changes showed a more categorical effect when supported by phrasal prosodic context, as compared to the same prominence manipulation in isolated words, we would have evidence that a more abstract prosodic information is being represented in the former case, but less so in the latter. This sort of asymmetry would support the architecture of MAPP in showing that prominence processing needn't necessarily incorporate prosodic analysis, though it will when phrasal prosodic context supports the integration of phonetic prominence cues.

5.3.2 Additive and conflicting prominence cues

Throughout the experiments in this dissertation prominence was manipulated as the presence or absence of a prominence-lending contextual cue, such that a target was cued as prominent, or not. We have seen that the presence of vowel-initial glottalization cues prominence to listeners, as do changes in contextual duration, f0, and amplitude signaling phrasal prosodic structure, but it's an open question how different prominence-lending cues might

interact when manipulated orthogonally. More generally, because prominence perception incorporates many pieces of information (Baumann & Cangemi, 2020; Baumann & Winter, 2018; Cole et al., 2019), it may be fruitful to test how various cues combine or compete in the domain of segmental processing. Imagine an experiment in which vowel-initial glottalization and phrasal prosodic prominence are manipulated orthogonally. In seeing if glottalization and phrasal prominence showed an additive effect in shifting listeners' perception, we could test their relative importance, and possible interactivity. For example, perhaps glottalization exerts a larger influence when a target is phrasally prominent (or perhaps, the two conflicting cues cancel one another out). Testing how orthogonally manipulated cues combine online would further replicate and extend the present findings to show how various prominence-lending properties mediate segmental processing both pre- and post-lexically. Testing when in processing these additive and competitive effects occurred would further inform us about the stage of prominence processing (pre- or post-lexical) in which listeners are integrating different cues. Different sorts of cues may pattern differently in this regard as well.

In this vein, we could also test the impact of orthogonally manipulated cues on the speed and/or accuracy of recognition. When cues come together to signal prominence (whether at a local/segmental level, or phrasally), they are predicted to facilitate the speed at which segmental material or words are categorized/recognized. Conflicting prominence cues might be expected to slow speech recognition, especially following MAPP, where they would feed forward to a parsed prosodic structure, which itself influences lexical competition. Anything that disrupts or slows the computation of prosodic structure is accordingly predicted to slow processing. As such, testing the extent to which conflicting prominence cues slow word recognition would be a useful test of the model and extension of the present results (cf. Braun, Dainora, & Ernestus, 2011; Nakai & Turk, 2011; Rysling et al., 2020).[6]

[6]Braun et al. (2011) found that an unfamiliar or unnatural intonation contour slowed word recognition, which could be taken as evidence in support of the MAPP proposal, where difficulty in computing prosody (including prominence cued by pitch) leads to slowed lexical processing.

5.3.3 Relation to boundary processing

As described above, there is empirical evidence that listeners' processing of prosodic boundary information is delayed (Kim, Mitterer, & Cho, 2018; Mitterer et al., 2019), compatible with the prosodic analysis model and the later stage integration of prosody in lexical competition as encoded in MAPP. MAPP predicts that localized/phonetic boundary cues should *not* influence processing at the same pre-lexical stage as prominence, given that boundary information is assumed to be linked more tightly to the computation of prosodic structure (i.e. listeners only have access to information about prosodic boundaries after prosodic analysis has begun). We could test this claim of MAPP by orthogonally manipulating prominence and boundary information, in a case where they would both be predicted to impact segmental perception (e.g., in VOT perception; Cole et al., 2007; Kim, Mitterer, & Cho, 2018). MAPP predicts that the influence of prominence will precede the influence of boundary. If this were indeed observed, it would lend strong support to the idea that boundary processing and prominence processing are different, in that only (phonetic) prominence processing is pre-lexical. It would also support the original prosodic analysis model of Cho et al. (2007). On the other hand, a clearly pre-lexical influence of boundary information, e.g., one that mirrored the influence of prominence information would be at odds with MAPP, and more generally with the findings of Kim, Mitterer, and Cho (2018) and Mitterer et al. (2019). As such, this sort of orthogonal prominence/boundary manipulation is seen as being a very useful test of the present proposal.

5.3.4 Integration of different prominence cues and modalities

The experiments in this dissertation limited themselves to manipulating information contained in the speech signal. However, it is well established that other signal-extrinsic pieces of information lend prominence in speech perception. A fruitful line of research would accordingly be testing if and how these other prominence-lending cues mediate segmental processing.

For example, Bishop (2012) implemented a prominence rating experiment in which listen-

ers heard sentences such as "I bought a motorcycle" preceded by written questions that set up different information structural expectations. When a preceding written context implied narrow focus on the object ("what did you buy?"), listeners perceived the object "motorcycle" as more prominent, in comparison to a broad focus context ("what happened?"). This was notably across conditions in which the actual acoustic stimulus was identical, showing clearly that information structure generated from the read context impacted perceived prominence. One promising line of research raised by the present results would accordingly be to test if prominence manipulations of this sort mediate segmental perception. If yes, we would have clear evidence for the involvement of signal-independent factors in segmental and prominence processing. Under the assumption that these sorts of information-structural manipulations feed into the computation of abstract prosodic structure (as shown in Figure 5.1), we could reconcile such a finding with the MAPP model in hypothesizing they involve only later stage modulation of lexical competition (resulting from an abstract prosodic parse). As such, we can predict they would show a strictly delayed influence in processing. Unlike Experiment 2, an earlier influence (due to phonetic prominence in the signal) would be absent, making this a case where we might be able to factor out early effects of prominence in processing. In that sense, this sort of extension could directly test MAPP's claim that abstract prominence information only enters into a later stage of processing. Disentangling early and later stage prominence effects in this way would help confirm the claims forwarded on the basis of the experiments in this dissertation, and open a wide range of questions about how much, and what sort of, top down information influences prominence perception as it pertains to segmental and lexical processing (see also Baumann & Winter, 2018; Cole, Mo, & Hasegawa-Johnson, 2010; Mo, 2011).

As a more general extension, it is also known that prosodic prominence is conveyed by hand movements, beat gestures, and facial expressions (e.g., Krivokapić, 2014; Swerts & Krahmer, 2008). Bosker and Peeters (2020) recently found that audiovisual presentation of speech with beat gestures influenced perception of metrical prominence, and even vowel duration as cue to phonemic vowel length. When a beat gesture co-occurred with a syllable, listeners were more likely to perceive that syllable as lexically stressed. Moreover, when a

beat gesture co-occurred with a syllable from an F2 continuum that cued a phonemically long or short Dutch vowel (cued by formant differences in addition to duration), listeners were more likely to categorize that vowel as phonemically short. The authors interpreted this as originating from listeners' perception of prominence on that vowel (due to a co-occurrent beat gesture), which led to an expectation of the vowel being relatively long. An ambiguous vowel was thus perceived as phonemically short in relation to the prominence-lending visual information, leading to more short vowel responses. This result shows just how important visual prominence perception may be in segmental processing. An obvious extension of Bosker and Peeters (2020) would be to explore if similar audiovisual integration obtained in the perception of vowel contrasts of the sort tested in this dissertation, that is, if beat gestures led to perceptual sonority expansion and hyperarticulation effects. Further observing the timecourse of audiovisual integration online would enrich our understanding of what constitutes prominence to listeners, and how it is brought to bear on segmental processing online.

5.3.5 Cross-linguistic prominence and segmental perception

Finally, another way in which these results can be extended more generally is in the investigation of how prominence strengthening in various languages is exploited perceptually by listeners. Languages will differ both in their phonological prosodic organization, and the way in which various aspects of prosody are signaled phonetically. Testing if speakers of different languages use prominence to a different extent, or in different ways, in segmental processing would help inform us both about cross-linguistic prosodic differences, and how they enter into speech recognition (in the vein of e.g. Cutler, Mehler, Norris, & Segui, 1986; Cutler & Otake, 1994).

Consider again the case of Tongan (Garellek & White, 2015) discussed in Chapter 4. Tongan showed uniform sonority expansion in terms of F1 for its five vowel system /i,e,a,o,u/. This stands in contrast to findings for American English for /i/ in particular (Cho, 2005; Kent & Netsell, 1971), and is the opposite of the perception results from Experiment 5 where

acoustically less sonorous versions of /i/ and /ɪ/ were expected in prominent contexts. Given that the perception results in Experiment 5 don't match directly with the patterns seen in the production literature for American English, we might hypothesize a general expectation of acoustic dispersion (paradigmatic enhancement) for high vowels. In the Tongan case, however, /i/ definitively undergoes sonority expansion, leaving it an open question how Tongan listeners would expect it to be strengthened phonetically when prominent. If Tongan listeners follow the pattern in their language they will expect a more sonorous /i/ under prominence. If they follow a general expectation of acoustic dispersion (as suggested by Experiment 5), a less sonorous /i/ will be expected.

At a more basic level, the role of prominence in segmental perception will likely vary from language to language, based on the extent to which prominence causes certain phonetic properties to vary, e.g., the extent to which non-prominent vowels are reduced (Delattre, 1969). For example, at the level of lexical prominence, it has been suggested that Spanish (among other languages) does not show substantial variation in vowel quality as a function of lexical stress, particularly in comparison to languages like English or Dutch (Delattre, 1969; Rietveld & Koopmans-van Beinum, 1987). Based on this pattern we could predict that variation in lexical prominence would engender an expectation of reduction in the latter case but not in the former. We can also predict these effects would go the other way in that vowel quality would contribute to prominence perception. For example, in a language with substantial vowel reduction, non-reduced spectral structure should cue prominence (Mo et al., 2009; Rietveld & Koopmans-van Beinum, 1987), but analogous variation in spectra might not inform prominence perception in a language like Spanish. Understanding how language-specific prominence effects fit into pre- and post-lexical prominence perception in a model like MAPP will necessitate a solid understanding of prosodic structure in a given language, and its phonetic encoding.

More broadly, developing a theory of how a language's segmental inventory relates to prominence strengthening effects in perception would help us understand the mapping between a segment's features and how that segment is expected to be strengthened under prominence. The asymmetries between high and non-high vowels seen in this dissertation

show that consideration of intrinsic vowel features is necessary, but the extent to which these effects are linked to the vowel inventory (more generally, segmental inventory) of a language is unclear. Based on the present results, one hypothesis is that paradigmatic enhancement effects for vowels will be stronger in languages with a crowded vowel space (cf. American English versus Tongan). In other words, listeners will prioritize perceptual dispersion in a vowel space under prominence, when a lack of dispersion would cause increased overlap in vowel categories (conceptually related to dispersion theory; Becker-Kristal, 2010; Flemming, 1996; Liljencrants & Lindblom, 1972). Of course, this cannot be the whole story, given the importance of sonority expansion shown in this dissertation. Testing the relative importance of these competing influences in speech perception, as done in speech production (Cho, 2005; de Jong, 1995) would be informative, particularly in testing languages with segmental inventories that varied in properties of interest.

More generally, explicating the role of segmental phonology (e.g., the prevalence of vowel reduction) for perceptual prominence strengthening patterns will help us understand how the two systems are related for both speakers and listeners.

5.4 Concluding remarks

To summarize, this dissertation has shown that prosodic prominence impacts segmental processing in various ways. We have seen that listeners adjust their perception of vowel contrasts on the basis of contextual prominence, where prominence-lending context engenders an expectation that a vowel will be realized in a phonetically strengthened manner. We have also seen that what counts as "strengthened" to listeners depends on vowel-intrinsic features. In addition to prominence marked by contextual changes in pitch and duration which cued phrasal/phonological organization, we have seen that glottalization in vowel-initial words also impacts vowel perception, lining up with its purported function as a prominence marker.

In terms of processing, we have seen that prominence effects are incorporated pre-lexically in listeners' mapping of cues to a segmental category. At the same time, phrasal prominence-lending context shows an overall delayed effect which is assumed to arise from post-lexical

integration of phrasal/phonological prominence information with activated lexical hypotheses in lexical competition. We have also seen that prominence effects, when cued in a strictly local fashion by glottalization, can be purely pre-lexical, showing the same contribution to lexical activation as vowel-intrinsic formant cues. Together, these findings show that prominence information is important throughout the process of speech recognition, and as such should be considered in models of prosodic and segmental processing such as the prosodic analysis model, and models of spoken word recognition more generally. To account for the present findings, the MAPP (Multi-stage Assessment of Prominence in Processing) model is proposed. MAPP offers a basic architecture compatible with the data, and provides a framework for further testing the claims made throughout this dissertation. The present findings are only the first step to a full theory of the role that prominence plays in segmental processing and spoken word recognition, and a theory of how these effects relate to prominence processing in other domains. Future research will help test and refine the MAPP proposal, and better our understanding of segment and prosody in speech processing more generally.

APPENDIX A

Appendix: Glottalization and spectral contrast

A.1 Experiment 7

As discussed in Chapter 3, changes in temporal and spectral context introduced by glottalization might exert an influence on listeners' perception of vowels based on spectral contrast, which could be confound in certain cases. Experiment 7 is included here to illustrate one such case, where spectral contrast (or, compensation for vowel-to vowel coarticulation) offers a possible explanation for effects that result from the manipulation of glottalization. In this experiment, listeners categorized a stimulus as "the ebb" or "the ab" with "the" pronounced as [ðiˈ], unlike Experiment 3 where the vowel [ə] was used in the precursor. This Experiment also differs from Experiment 3 in that the stimuli were just a two-word sequence, as compared to "say the ebb/ab now".

As in Experiment 3, the crucial manipulation was whether a glottal stop intervened preceding the target vowel. In the glottal stop condition, the target and precursor are separated by [ʔ], while in the no glottal stop condition, F1 and F2 transition continuously from precursor to target. Note that [i] has lower F1 and higher F2, relative to all steps on the continuum. We could therefore conceptualize an [i] precursor as being more *peripheral* to all continuum steps in F1/F2 terms. This can be compared to Experiment 3, where a fairly low [ə] (that is, an [ə] with F1 which was manipulated to be relatively high) was overall less peripheral in F1 and F2 than the steps on the continuum. To illustrate this, formant tracks from the stimuli in Experiment 7 and Experiment 3 (also Experiment 4) are shown in Figure A.1. Note how the relationship between the precursor formants and the target continuum is generally reversed across experiments. The relevance of this difference will be discussed.

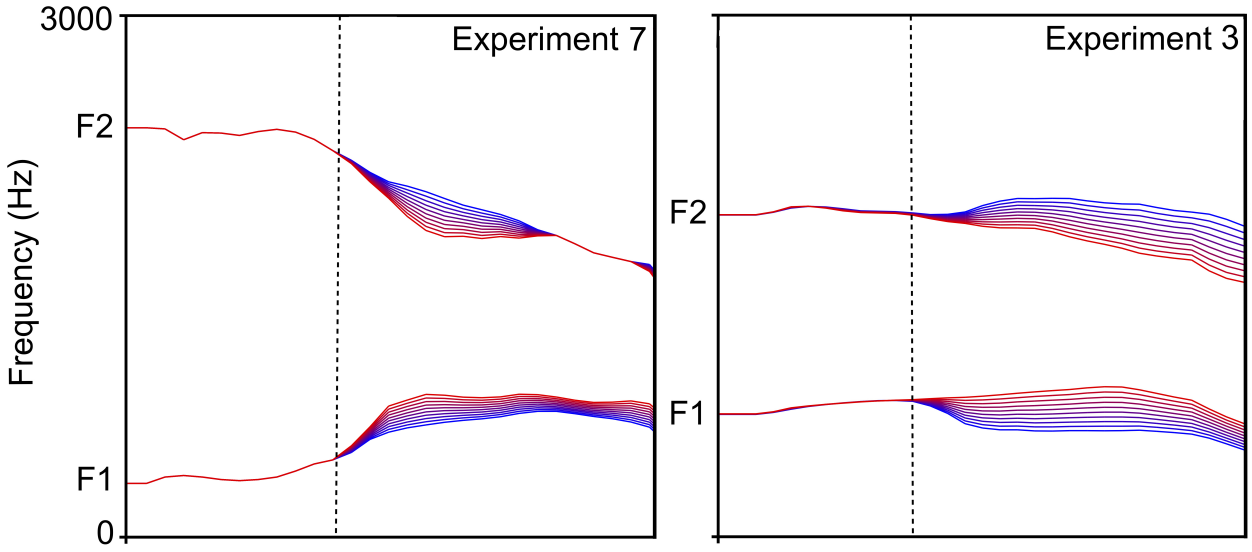

Figure A.1: Formant tracks comparing the continua in Experiment 7 and Experiment 3, with Frequency (0-3000 Hz) on the y axis, and time on the x axis. F1 and F2 are indexed to the left of each plot. The precursor vowel ([i] in Experiment 7, [ə] in Experiment 3) is separated from the target continuum by a dashed vertical line, indicating where the glottal stop intervened. The target continua are arrayed such that the F1 and F2 values for the /æ/ endpoint are the innermost red lines, and the F1 and F2 values for the /ɛ/ endpoint are the outermost blue lines, as in Figure 2.2 and 3.2.

A.1.1 Materials

The method of stimulus manipulation in Experiment 7 was the same as in Experiment 3. The starting point for stimulus manipulation was a production of the two word sequence "the ebb" with the word "the" pronounced as [ði]. The resynthesis altered only F1 and F2 in the target word leaving the precursor [ði] unaltered. The result was accordingly a [ðiɛb] to [ðiæb] continuum, as shown in Figure A.1, with continuous formant transitions from the precursor vowel to the target. This constitutes what will subsequently be referred to as the no glottal stop condition, that is, where no glottal stop preceded the target sound in the hiatus environment. Following this, a glottal stop was spliced to intervene between the precursor and target vowel, using the same method as in Experiment 3, and resulting in

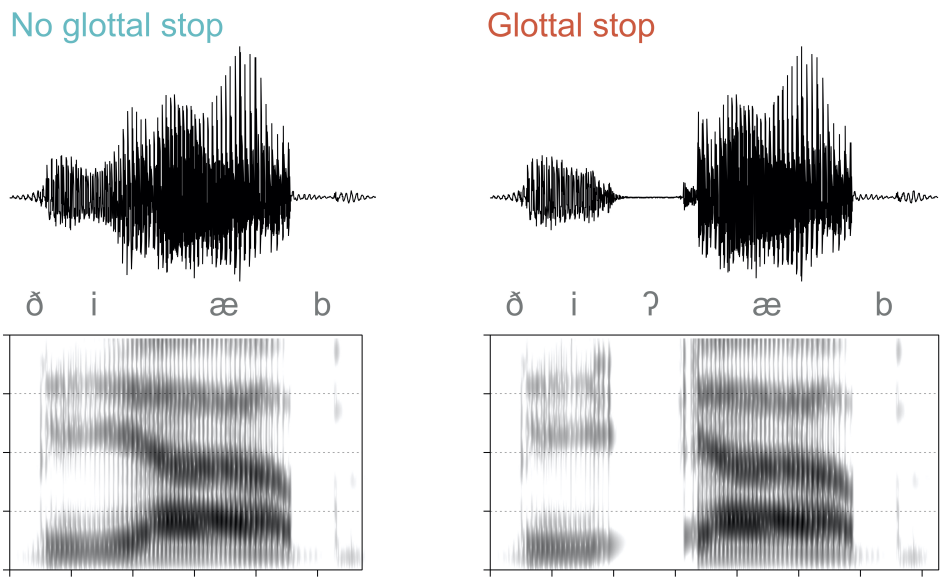

Figure A.2: Waveforms and spectrograms of the Experiment 7 stimuli. A segmental transcription is given in IPA above the spectrograms. In the spectrogram, ticks on the y axis indicate 1000Hz, for a frequency range of 0-4000Hz. Ticks on the x axis are placed at every 100 ms. The target word shown in the figure is from the /æ/ endpoint of the continuum.

a [ðiʔɛb] to [ðiʔæb] continuum, which constituted the glottal stop condition. Examples of these two conditions are shown in Figure A.2.

A.1.2 Participants and procedure

30 participants were recruited for Experiment 7 from the same population as previous experiments. They completed the experiment in the lab (not remotely). The procedure, number of trials, etc. were identical to Experiment 3: it was a simple 2AFC task in which participants heard a stimulus and categorized it as one of two words, "ebb" or "ab" (see Section 3.3.2).

A.1.3 Results and discussion

The statistical assessment of the Experiment 7 results was the same as all previous categorization data. In coding the dependent variable, an "ebb" response was mapped to 1, and an

Table A.1: Model output for Experiment 7.

	Estimate	Est. Error	L-95% CI	U-95%CI	credible?
intercept	-0.05	0.16	-0.37	0.28	
glottal stop	1.27	0.12	1.03	1.52	✓
continuum	-2.26	0.20	-2.65	-1.88	✓
glottal stop:continuum	0.06	0.06	-0.06	0.17	

"ab" response was mapped to 0. In contrast-coding the glottal stop manipulation, a glottal stop was mapped to 0.5, and no glottal stop was mapped to -0.5. The model output is shown in Table A.1 and categorization responses are plotted in Figure A.3.

As would be expected, changing formant values along the continuum showed a credible effect, whereby increasing continuum step (becoming less "ebb"-like), decreased "ebb" responses (β=-2.26, 95%CI =[-2.65,-1.88]). A robust effect of preceding glottal stop was also observed, whereby the presence of a preceding glottal stop credibly increased listeners' "ebb" responses (β=1.27, 95%CI =[1.03,1.52]). The directionality of this effect is notably the same as the effect observed for Experiment 1 and 3, which could be taken to support the prominence predictions forwarded throughout this dissertation.

Consider, however, how the results from Experiment 7 might be explained by spectral contrast, as described in Section 3.3.1. Contrast effects in this experiment would predict that target F1 and F2 would be perceived by listeners as relatively centralized in the spectrum (that is, F1 perceived as higher, F2 perceived as lower). In other words, a vowel following the high front vowel precursor would be perceived as relatively low and back (here, more "ab"-like). We can also frame this in terms of coarticulation: /æ/ that is coarticulated with /i/ will become acoustically more like /ɛ/. The presence of a preceding /i/ would therefore generally be expected to increase "ab" responses (or, decrease "ebb" responses) in the present stimuli (as compared to, e.g., a preceding low vowel). Now consider the temporal interruption (approximately 100 ms), and discontinuity in formant trajectories, introduced by an intervening glottal stop (in the glottal stop condition). In temporally separating the

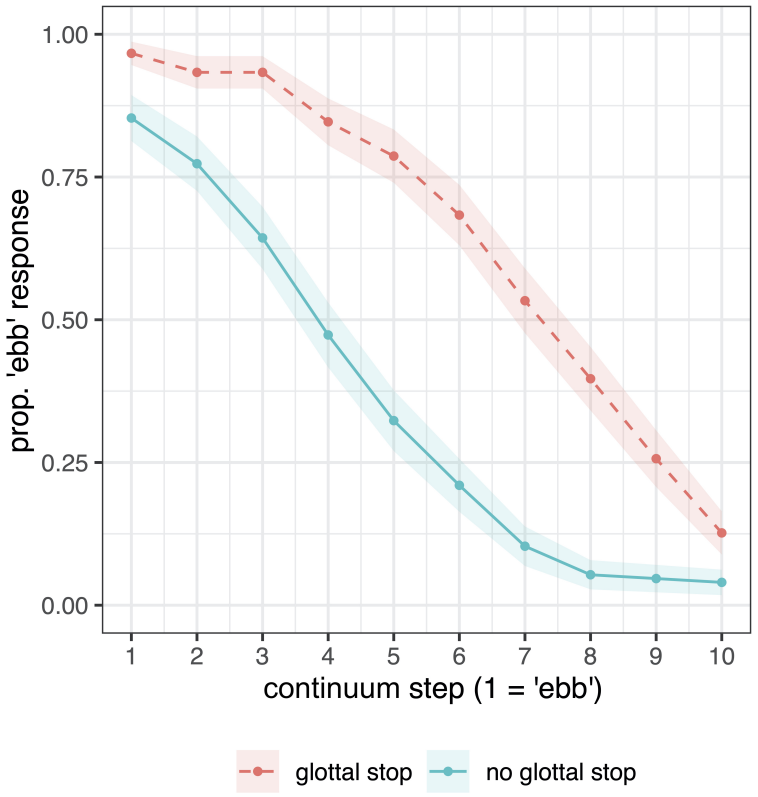

Figure A.3: Categorization responses in Experiment 7, with the proportion of "ebb" responses plotted on the y axis, split by prominence condition and continuum step, where step 1 is the /ɛ/ endpoint of the continuum. Shading around each line shows 95% CI.

target and the precursor, a glottal stop should reduce the strength of contrast effects in the no glottal stop condition, that is, where the no glottal stop condition should show decreased "ebb" responses due to spectral contrast, the glottal stop condition should reduce this effect, showing relatively increased "ebb" responses. Framing this effect in coarticulatory terms, a glottal stop might reduce perceived coarticulation between two adjacent vowels and therefore decrease listeners' attribution of target formant structure to the precursor vowel, i.e. an [iæ] sequence would show a stronger coarticulatory influence on the second vowel, as compared to an [iʔæ] sequence, such that more "ab" responses are obtained for [iæ].

This pattern of results is indeed what was found in Experiment 7, leaving open the possibility that spectral contrast effects might be partially, or fully, responsible for the observed

shift in categorization. The results of Experiment 7 therefore cannot provide definitive evidence that a glottal stop cues prominence to listeners, shifting their perception of vowel contrasts (see Mitterer et al., 2016 and Steffman, 2019a for an analogous discussion of prosodic boundary effects and durational contrast).

The crucial difference between Experiment 7 and Experiment 3 is the relationship between spectral energy in the precursor and target, as shown in Figure A.1. With more peripheral F1 and F2 in the precursor in Experiment 7, predicted contrast effects are a confound. In Experiment 3, with less peripheral F1 and F2 in the precursor, contrast effects predict the opposite of the glottalization-as-prominence account. The fact that the same overall result obtains in both experiments (increased "ebb" responses in the glottal stop condition), offers support for the idea that a glottal stop does indeed cue prominence, though we would not be able to make this claim on the basis of Experiment 7 alone. Also of note in these results, the magnitude of the effect is larger in Experiment 7 as compared to Experiment 3. This suggests the possibility that spectral contrast effects are playing a role such that listeners' perception of the target is primarily impacted by glottalization as prominence, but can further be shifted by contrast effects, or in the case of Experiment 3, diminished (though clearly still present) by competing spectral contrast.

As Experiment 7 illustrates, controlling for other possible influences (as done in Experiment 3) is an important step in exploring prosodic effects that change spectral and durational context. This presents a similar line of argument to points raised by Mitterer et al. (2016) and Steffman (2019a, 2019b).

APPENDIX B

Appendix: GAMM model outputs

Table B.1: Model output for the GAMM used in Experiment 2, with parametric terms shown above and smooth terms shown below.

Parametric terms	Estimate	Est. Error	t-value	p-value
intercept	0.24	0.16	1.50	0.14
continuum	-1.63	0.09	-18.04	< 0.001
prominence	0.45	0.23	1.91	0.06
Smooth terms	**edf**	**ref df**	**F-value**	**p-value**
te(time, continuum; condition = NPA)	17.09	19.71	38.27	< 0.001
te(time, continuum; condition = post-focus)	8.99	9.52	66.32	< 0.001
s(time, participant; condition = NPA)	228.11	323.00	3.91	< 0.001
s(time, participant; condition = post-focus)	231.32	323.00	4.97	< 0.001

Table B.2: Model output for the GAMM used in Experiment 4, with parametric terms shown above and smooth terms shown below. "GS" refers to "glottal stop".

Parametric terms	Estimate	Est. Error	t-value	p-value
intercept	0.37	0.08	4.39	$<$0.001
continuum	-1.01	0.10	-9.86	< 0.001
glottal stop	0.52	0.12	4.45	< 0.001
Smooth terms	edf	ref df	F-value	p-value
te(time, continuum; condition = GS)	21.85	22.48	333.36	< 0.001
te(time, continuum; condition = no GS)	22.00	23.05	282.845	< 0.001
s(time, participant; condition = GS)	268.60	358.00	9.38	< 0.001
s(time, participant; condition = no GS)	284.13	358.00	10.24	< 0.001

References

Abramson, A. S. (1976). Laryngeal timing in consonant distinctions. *Haskins Laboratory Status Report on Speech Research, SR-47*, 105–112.

Abramson, A. S., & Whalen, D. H. (2017). Voice Onset Time (VOT) at 50: Theoretical and practical issues in measuring voicing distinctions. *Journal of Phonetics*, *63*, 75–86.

Arndt-Lappe, S., & Ernestus, M. (2020). Morpho-phonological alternations: The role of lexical storage. In *Word Knowledge and Word Usage* (pp. 191–227). De Gruyter Mouton.

Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge University Press.

Baayen, R. H., van Rij, J., de Cat, C., & Wood, S. (2018). Autocorrelated errors in experimental data in the language sciences: Some solutions offered by generalized additive mixed models. In *Mixed-Effects Regression Models in Linguistics* (pp. 49–69). Springer.

Babel, M. (2009). *Phonetic and social selectivity in speech accommodation* (Unpublished doctoral dissertation). University of California, Berkeley.

Baese-Berk, M. M., Heffner, C. C., Dilley, L. C., Pitt, M. A., Morrill, T. H., & McAuley, J. D. (2014). Long-term temporal tracking of speech rate affects spoken-word recognition. *Psychological Science*, *25*(8), 1546–1553.

Barr, D. J. (2008). Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of Memory and Language*, *59*(4), 457–474.

Baumann, S., & Cangemi, F. (2020). Integrating phonetics and phonology in the study of linguistic prominence. *Journal of Phonetics*, *81*, 100993.

Baumann, S., & Schumacher, P. B. (2020). The incremental processing of focus, givenness and prosodic prominence. *Glossa: a journal of general linguistics*, *5*(1).

Baumann, S., & Winter, B. (2018). What makes a word prominent? Predicting untrained German listeners' perceptual judgments. *Journal of Phonetics*, *70*, 20–38.

Becker-Kristal, R. (2010). *Acoustic typology of vowel inventories and dispersion theory:*

Insights from a large cross-linguistic corpus (Unpublished doctoral dissertation). University of California, Los Angeles.

Beckman, J. N. (1998). *Positional faithfulness* (Unpublished doctoral dissertation). University of Massachusetts Amherst.

Beckman, M. E. (1996). The Parsing of Prosody. *Language and Cognitive Processes*, *11*(1-2), 17–68.

Beckman, M. E., & Ayers, G. (1997). Guidelines for ToBI labelling. *The OSU Research Foundation*, *3*.

Beckman, M. E., Edwards, J., & Fletcher, J. (1992). Prosodic structure and tempo in a sonority model of articulatory dynamics. In G. J. Docherty & D. R. Ladd (Eds.), *Gesture, Segment, Prosody* (pp. 68–89). Cambridge University Press.

Beckman, M. E., & Pierrehumbert, J. B. (1986). Intonational structure in Japanese and English. *Phonology*, *3*, 255–309.

Beddor, P. S., Harnsberger, J. D., & Lindemann, S. (2002). Language-specific patterns of vowel-to-vowel coarticulation: Acoustic structures and their perceptual correlates. *Journal of Phonetics*, *30*(4), 591–627.

Berkovits, R. (1993). Utterance-final lengthening and the duration of final-stop closures. *Journal of Phonetics*, *21*(4), 479–489.

Bigand, E., & Pineau, M. (1997). Global context effects on musical expectancy. *Perception & Psychophysics*, *59*(7), 1098–1107.

Bishop, J. (2012). Information structural expectations in the perception of prosodic prominence. In *Prosody and Meaning* (pp. 239–269). Walter de Gruyter.

Bishop, J. (2017). Focus projection and prenuclear accents: Evidence from lexical processing. *Language, Cognition and Neuroscience*, *32*(2), 236–253.

Bishop, J., Kuo, G., & Kim, B. (2020). Phonology, phonetics, and signal-extrinsic factors in the perception of prosodic prominence: Evidence from rapid prosody transcription. *Journal of Phonetics*, *82*, 100977.

Blouin, D. C., & Riopelle, A. J. (2005). On confidence intervals for within-subjects designs. *Psychological Methods*, *10*(4), 397–412.

Boersma, P., & Weenink, D. (2020). *Praat: doing phonetics by computer (version 6.1.09).* Retrieved from `http://www.praat.org`

Bolinger, D. L. (1958). A theory of pitch accent in English. *Word*, *14*(2-3), 109–149.

Bolinger, D. L. (1961). Contrastive accent and contrastive stress. *Language*, *37*(1), 83–96.

Bosker, H. R. (2017). Accounting for rate-dependent category boundary shifts in speech perception. *Attention, Perception, & Psychophysics*, *79*(1), 333–343.

Bosker, H. R., & Peeters, D. (2020). Beat gestures influence which speech sounds you hear. *bioRxiv*.

Bosker, H. R., Reinisch, E., & Sjerps, M. J. (2017). Cognitive load makes speech sound fast, but does not modulate acoustic context effects. *Journal of Memory and Language*, *94*, 166–176.

Brand, S., & Ernestus, M. (2018). Listeners' processing of a given reduced word pronunciation variant directly reflects their exposure to this variant: Evidence from native listeners and learners of French. *Quarterly Journal of Experimental Psychology*, *71*(5), 1240–1259.

Braun, B., Dainora, A., & Ernestus, M. (2011). An unfamiliar intonation contour slows down online speech comprehension. *Language and Cognitive Processes*, *26*(3), 350–375.

Bregman, A. S. (1994). *Auditory Scene Analysis: The Perceptual Organization of Sound.* Massachusetts Institute of Technology Press.

Brigner, W. L. (1988). Perceived duration as a function of pitch. *Perceptual and Motor Skills*, *67*(1), 301–302.

Brown, M., Salverda, A. P., Dilley, L., & Tanenhaus, M. K. (2015). Metrical expectations from preceding prosody influence perception of lexical stress. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(2), 306.

Brunner, J., & Zygis, M. (2011). Why do glottal stops and low vowels like each other? In *Proceedings of the 17th International Congress of Phonetic Sciences* (pp. 376–379).

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

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28.

Byrd, D., & Saltzman, E. (2003). The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening. *Journal of Phonetics*, *31*(2), 149–180.

Cairns, H. S., & Hsu, J. R. (1980). Effects of prior contect on lexical access during sentence comprehension: A replication and reinterpretation. *Journal of Psycholinguistic Research*, *9*(4), 319–326.

Calhoun, S. (2007). *Information structure and the prosodic structure of English: A probabilistic relationship* (Unpublished doctoral dissertation). University of Edinburgh.

Cangemi, F., & Grice, M. (2016). The importance of a distributional approach to categoriality in autosegmental-metrical accounts of intonation. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *7*(1).

Carlson, K., Clifton, C., & Frazier, L. (2001). Prosodic boundaries in adjunct attachment. *Journal of Memory and Language*, *45*(1), 58–81.

Chen, M. (1970). Vowel length variation as a function of the voicing of the consonant environment. *Phonetica*, *22*(3), 129–159.

Cho, T. (2004). Prosodically conditioned strengthening and vowel-to-vowel coarticulation in English. *Journal of Phonetics*, *32*(2), 141–176.

Cho, T. (2005). Prosodic strengthening and featural enhancement: Evidence from acoustic and articulatory realizations of /ɑ, i/ in English. *The Journal of the Acoustical Society of America*, *117*(6), 3867–3878.

Cho, T. (2015). Language Effects on Timing at the Segmental and Suprasegmental Levels. In M. A. Redford (Ed.), *The Handbook of Speech Production* (pp. 505–529). John Wiley & Sons, Inc.

Cho, T. (2016). Prosodic Boundary Strengthening in the Phonetics–Prosody Interface. *Language and Linguistics Compass*, *10*(3), 120–141.

Cho, T., & Jun, S.-A. (2000). Domain-initial strengthening as enhancement of laryngeal features: Aerodynamic evidence from Korean. *UCLA Working Papers in Phonetics*, 57–70.

Cho, T., & Keating, P. (2001). Articulatory and acoustic studies on domain-initial strengthening in Korean. *Journal of Phonetics*, *29*(2), 155–190.

Cho, T., & Keating, P. (2009). Effects of initial position versus prominence in English. *Journal of Phonetics*, *37*(4), 466–485.

Cho, T., Kim, D., & Kim, S. (2017). Prosodically-conditioned fine-tuning of coarticulatory vowel nasalization in English. *Journal of Phonetics*, *64*, 71–89.

Cho, T., Lee, Y., & Kim, S. (2014). Prosodic strengthening on the/s/-stop cluster and the phonetic implementation of an allophonic rule in English. *Journal of Phonetics*, *46*, 128–146.

Cho, T., & McQueen, J. M. (2005). Prosodic influences on consonant production in Dutch: Effects of prosodic boundaries, phrasal accent and lexical stress. *Journal of Phonetics*, *33*(2), 121–157.

Cho, T., McQueen, J. M., & Cox, E. A. (2007). Prosodically driven phonetic detail in speech processing: The case of domain-initial strengthening in English. *Journal of Phonetics*, *35*(2), 210–243.

Chodroff, E., & Wilson, C. (2019). Acoustic–phonetic and auditory mechanisms of adaptation in the perception of sibilant fricatives. *Attention, Perception, & Psychophysics*, 1–22.

Choi, J., Kim, S., & Cho, T. (2020). An apparent-time study of an ongoing sound change in Seoul Korean: A prosodic account. *Plos one*, *15*(10), e0240682.

Chomsky, N., & Halle, M. (1968). *The Sound Pattern of English*. Harper & Row New York.

Chong, J., & Garellek, M. (2018). Online perception of glottalized coda stops in American English. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*.

Christophe, A., Peperkamp, S., Pallier, C., Block, E., & Mehler, J. (2004). Phonological phrase boundaries constrain lexical access I. Adult data. *Journal of Memory and Language*, *51*(4), 523–547.

Chuang, C.-K., & Wang, W. S.-Y. (1976). Influence of vowel height, intensity, and temporal order on pitch perception. *The Journal of the Acoustical Society of America*, *60*(S1), S92–S92.

Chuang, Y.-Y., & Fon, J. (2010). The effect of prosodic prominence on the realizations of voiceless dental and retroflex sibilants in Taiwan Mandarin spontaneous speech. In *Proceedings of the 5th International Conference on Speech Prosody.*

Clopper, C. G., Pisoni, D. B., & de Jong, K. (2005). Acoustic characteristics of the vowel systems of six regional varieties of American English. *The Journal of the Acoustical Society of America*, *118*(3), 1661–1676.

Coady, J. A., Kluender, K. R., & Rhode, W. S. (2003). Effects of contrast between onsets of speech and other complex spectra. *The Journal of the Acoustical Society of America*, *114*(4), 2225–2235.

Cole, J., Hualde, J. I., Smith, C. L., Eager, C., Mahrt, T., & de Souza, R. N. (2019). Sound, structure and meaning: The bases of prominence ratings in English, French and Spanish. *Journal of Phonetics*, *75*, 113–147.

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*(2), 180–209.

Cole, J., Linebaugh, G., Munson, C., & McMurray, B. (2010). Unmasking the acoustic effects of vowel-to-vowel coarticulation: A statistical modeling approach. *Journal of Phonetics*, *38*(2), 167–184.

Cole, J., Mo, Y., & Hasegawa-Johnson, M. (2010). Signal-based and expectation-based factors in the perception of prosodic prominence. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *1*(2), 425–452.

Cole, J., & Shattuck-Hufnagel, S. (2016). New methods for prosodic transcription: Capturing variability as a source of information. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *7*(1).

Cole, J., Shattuck-Hufnagel, S., & Mo, Y. (2010). Prosody production in spontaneous speech: Phonological encoding, phonetic variability, and the prosodic signature of individual speakers. *The Journal of the Acoustical Society of America*, *128*(4), 2429–2429.

Cole, R. A., & Jakimik, J. (1980). How are syllables used to recognize words? *The Journal of the Acoustical Society of America*, *67*(3), 965–970.

Connaghan, K. P., & Patel, R. (2017). The impact of contrastive stress on vowel acoustics and intelligibility in dysarthria. *Journal of Speech, Language, and Hearing Research*, *60*(1), 38–50.

Connine, C. M., Titone, D., & Wang, J. (1993). Auditory word recognition: Extrinsic and intrinsic effects of word frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(1), 81.

Cooper, N., Cutler, A., & Wales, R. (2002). Constraints of lexical stress on lexical access in English: Evidence from native and non-native listeners. *Language and Speech*, *45*(3), 207–228.

Cusack, R., Decks, J., Aikman, G., & Carlyon, R. P. (2004). Effects of location, frequency region, and time course of selective attention on auditory scene analysis. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(4), 643.

Cutler, A. (1976). Phoneme-monitoring reaction time as a function of preceding intonation contour. *Perception & Psychophysics*, *20*(1), 55–60.

Cutler, A. (2010). Abstraction-based efficiency in the lexicon. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *1*(2), 301–318.

Cutler, A., Dahan, D., & Van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech*, *40*(2), 141–201.

Cutler, A., Eisner, F., McQueen, J. M., & Norris, D. (2006). Coping with speaker-related variation via abstract phonemic categories. In *Proceedings of the 10th Conference on Laboratory Phonology* (pp. 31–32).

Cutler, A., & Foss, D. J. (1977). On the role of sentence stress in sentence processing. *Language and Speech*, *20*(1), 1–10.

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1986). The syllable's differing role in the segmentation of French and English. *Journal of Memory and Language*, *25*, 385–400.

Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(1), 113.

Cutler, A., & Otake, T. (1994). Mora or phoneme? Further evidence for language-specific

listening. *Journal of Memory and Language*, *33*(6), 824–844.

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*(5-6), 507–534.

Davidson, L. (2016). Variability in the implementation of voicing in American English obstruents. *Journal of Phonetics*, *54*, 35–50.

de Jong, K. (1991). *The oral articulation of English stress accent* (Unpublished doctoral dissertation). The Ohio State University.

de Jong, K. (1995). The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *The Journal of the Acoustical Society of America*, *97*(1), 491–504.

de Jong, K., Beckman, M. E., & Edwards, J. (1993). The interplay between prosodic structure and coarticulation. *Language and Speech*, *36*(2-3), 197–212.

Delattre, P. (1969). An acoustic and articulatory study of vowel reduction in four languages. *IRAL: International Review of Applied Linguistics in Language Teaching*, *7*(4), 295–325.

Diehl, R. L., Souther, A. F., & Convis, C. L. (1980). Conditions on rate normalization in speech perception. *Perception & Psychophysics*, *27*(5), 435–443.

Diehl, R. L., & Walsh, M. A. (1989). An auditory basis for the stimulus-length effect in the perception of stops and glides. *The Journal of the Acoustical Society of America*, *85*(5), 2154–2164.

Dilley, L., Mattys, S. L., & Vinke, L. (2010). Potent prosody: Comparing the effects of distal prosody, proximal prosody, and semantic context on word segmentation. *Journal of Memory and Language*, *63*(3), 274–294.

Dilley, L., Shattuck-Hufnagel, S., & Ostendorf, M. (1996). Glottalization of word-initial vowels as a function of prosodic structure. *Journal of Phonetics*, *24*(4), 423–444.

Dilley, L. C., & Heffner, C. (2013). The role of F0 alignment in distinguishing intonation categories: evidence from American English. *Journal of Speech Sciences*, *3*(1), 3–67.

Dilley, L. C., Ladd, D. R., & Schepman, A. (2005). Alignment of L and H in bitonal pitch

accents: testing two hypotheses. *Journal of Phonetics*, *33*(1), 115–119.

Edwards, J., & Beckman, M. E. (1988). Articulatory timing and the prosodic interpretation of syllable duration. *Phonetica*, *45*(2-4), 156–174.

Endress, A. D., & Hauser, M. D. (2010). Word segmentation with universal prosodic cues. *Cognitive Psychology*, *61*(2), 177–199.

Erickson, D. (2002). Articulation of extreme formant patterns for emphasized vowels. *Phonetica*, *59*(2-3), 134–149.

Ernestus, M. (2014). Acoustic reduction and the roles of abstractions and exemplars in speech processing. *Lingua*, *142*, 27–41.

Fant, G., & Kruckenberg, A. (1989). Preliminaries to the study of Swedish prose reading and reading style. *STL-QPSR*, *2*(1989), 1–83.

Fischer, B. (1992). Saccadic reaction time: Implications for reading, dyslexia, and visual cognition. In *Eye Movements and Visual Cognition* (pp. 31–45). Springer.

Flemming, E. (1996). Evidence for constraints on contrast: The dispersion theory of contrast. *UCLA Working Papers in Phonology*, *1*, 86–106.

Fletcher, J. (2004). An EMA/EPG study of vowel-to-vowel articulation across velars in Southern British English. *Clinical Linguistics & Phonetics*, *18*(6-8), 577–592.

Fletcher, J. (2010). The Prosody of Speech: Timing and Rhythm. In *The Handbook of Phonetic Sciences* (p. 521-602). John Wiley & Sons, Inc.

Fletcher, J., Stoakes, H., Loakes, D., & Singer, R. (2015). Accentual prominence and consonant lengthening and strengthening in Mawng. In *Proceedings of the 18th International Congress of Phonetic Sciences.*

Fougeron, C. (2001). Articulatory properties of initial segments in several prosodic constituents in French. *Journal of Phonetics*, *29*(2), 109–135.

Fougeron, C., & Keating, P. (1997). Articulatory strengthening at edges of prosodic domains. *Journal of the Acoustical Society of America*, *106*(6), 3728–3740.

Fowler, C. A. (1981). Production and perception of coarticulation among stressed and unstressed vowels. *Journal of Speech, Language, and Hearing Research*, *24*(1), 127–139.

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

Fowler, C. A., & Housum, J. (1987). Talkers' signaling of "new" and "old" words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language*, *26*(5), 489–504.

Friederici, A. D., Steinhauer, K., & Frisch, S. (1999). Lexical integration: Sequential effects of syntactic and semantic information. *Memory & Cognition*, *27*(3), 438–453.

Fujimura, O. (1990). Methods and goals of speech production research. *Language and Speech*, *33*(3), 195–258.

Garellek, M. (2011). The benefits of vowel laryngealization on the perception of coda stops in English. *UCLA Working Papers in Phonetics*, *109*, 31–39.

Garellek, M. (2013). *Production and perception of glottal stops* (Unpublished doctoral dissertation). University of California, Los Angeles.

Garellek, M. (2014). Voice quality strengthening and glottalization. *Journal of Phonetics*, *45*, 106–113.

Garellek, M. (2015). Perception of glottalization and phrase-final creak. *The Journal of the Acoustical Society of America*, *137*(2), 822–831.

Garellek, M., & Keating, P. (2011). The acoustic consequences of phonation and tone interactions in Jalapa Mazatec. *Journal of the International Phonetic Association*, 185–205.

Garellek, M., & Seyfarth, S. (2016). Acoustic Differences Between English /t/ Glottalization and Phrasal Creak. In *Proceedings of INTERSPEECH* (pp. 1054–1058).

Garellek, M., & White, J. (2015). Phonetics of Tongan stress. *Journal of the International Phonetic Association*, *45*(01), 13–34.

Georgeton, L., Antolík, T. K., & Fougeron, C. (2016). Effect of domain initial strengthening on vowel height and backness contrasts in French: Acoustic and ultrasound data. *Journal of Speech, Language, and Hearing Research*, *59*(6), S1575–S1586.

Gerfen, C., & Baker, K. (2005). The production and perception of laryngealized vowels in Coatzospan Mixtec. *Journal of Phonetics*, *33*(3), 311–334.

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

Gordon, M., & Ladefoged, P. (2001). Phonation types: a cross-linguistic overview. *Journal of Phonetics*, *29*(4), 383–406.

Green, K. P., Tomiak, G. R., & Kuhl, P. K. (1997). The encoding of rate and talker information during phonetic perception. *Perception & Psychophysics*, *59*(5), 675–692.

Grice, M., Ritter, S., Niemann, H., & Roettger, T. B. (2017). Integrating the discreteness and continuity of intonational categories. *Journal of Phonetics*, *64*, 90–107.

Guitard-Ivent, F., Chignoli, G., Fougeron, C., & Georgeton, L. (2019). Are IP initial vowels acoustically more distinct? Results from LDA and CNN. In *Proceedings of INTERSPEECH* (pp. 1746–1750).

Hagiwara, R. (1997). Dialect variation and formant frequency: The American English vowels revisited. *The Journal of the Acoustical Society of America*, *102*(1), 655–658.

Hanson, H. M., Stevens, K. N., Kuo, H.-K. J., Chen, M. Y., & Slifka, J. (2001). Towards models of phonation. *Journal of Phonetics*, *29*(4), 451–480.

Harrington, J., Kleber, F., & Reubold, U. (2013). The effect of prosodic weakening on the production and perception of trans-consonantal vowel coarticulation in German. *The Journal of the Acoustical Society of America*, *134*(1), 551–561.

Hawkins, S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics*, *31*(3-4), 373–405.

Hayes, B. (1995). *Metrical Stress Theory: Principles and Case Studies*. University of Chicago Press.

Heffner, C. C., Newman, R. S., & Idsardi, W. J. (2017). Support for context effects on segmentation and segments depends on the context. *Attention, Perception, & Psychophysics*, *79*(3), 964–988.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical society of America*, *97*(5), 3099–3111.

Hillenbrand, J. M., & Houde, R. A. (1996). Role of f0 and amplitude in the perception of

intervocalic glottal stops. *Journal of Speech, Language, and Hearing Research*, *39*(6), 1182–1190.

Hirano, M., Ohala, J., & Vennard, W. (1969). The function of laryngeal muscles in regulating fundamental frequency and intensity of phonation. *Journal of Speech and Hearing Research*, *12*(3), 616–628.

Hirschberg, J., & Pierrehumbert, J. (1986). The intonational structuring of discourse. In *Proceedings of the 24th Annual Meeting of the Association for Computational Linguistics* (pp. 136–144).

Holt, L. L. (2005). Temporally nonadjacent nonlinguistic sounds affect speech categorization. *Psychological Science*, *16*(4), 305–312.

Holt, L. L. (2006). The mean matters: Effects of statistically defined nonspeech spectral distributions on speech categorization. *The Journal of the Acoustical Society of America*, *120*(5), 2801–2817.

Holt, L. L., Lotto, A. J., & Kluender, K. R. (2000). Neighboring spectral content influences vowel identification. *The Journal of the Acoustical Society of America*, *108*(2), 710–722.

Houde, R. (1967). *A study of tongue body movement during selected speech sounds* (Unpublished doctoral dissertation). University of Michigan.

House, A. S. (1961). On vowel duration in English. *The Journal of the Acoustical Society of America*, *33*(9), 1174–1178.

Ito, K., & Speer, S. R. (2008). Anticipatory effects of intonation: Eye movements during instructed visual search. *Journal of Memory and Language*, *58*(2), 541–573.

Jones, M. R., & McAuley, J. D. (2005). Time judgments in global temporal contexts. *Perception & Psychophysics*, *67*(3), 398–417.

Jun, S.-A. (1996). *The phonetics and phonology of Korean prosody: Intonational phonology and prosodic structure*. Taylor & Francis.

Jun, S.-A. (1998). The accentual phrase in the Korean prosodic hierarchy. *Phonology*, 189–226.

Jun, S.-A. (2005). *Prosodic Typology: The Phonology of Intonation and Phrasing* (Vol. 1).

Oxford University Press.

Jun, S.-A. (2014). *Prosodic Typology II: The Phonology of Intonation and Phrasing* (Vol. 2). Oxford University Press.

Kang, Y. (2014). Voice Onset Time merger and development of tonal contrast in Seoul Korean stops: A corpus study. *Journal of Phonetics*, *45*, 76–90.

Katsika, A. (2016). The role of prominence in determining the scope of boundary-related lengthening in Greek. *Journal of Phonetics*, *55*, 149–181.

Katz, J., & Selkirk, E. (2011). Contrastive focus vs. discourse-new: Evidence from phonetic prominence in English. *Language*, 771–816.

Keating, P. (2006). Phonetic encoding of prosodic structure. In *Speech Production: Models, Phonetic Processes, and Techniques* (pp. 167–186). Psychology Press.

Keating, P., Cho, T., Fougeron, C., & Hsu, C.-S. (2004). Domain-initial articulatory strengthening in four languages. In *Phonetic interpretation: Papers in Laboratory Phonology VI* (pp. 143–161). Cambridge University Press.

Keating, P., & Shattuck-Hufnagel, S. (2002). A prosodic view of word form encoding for speech production. *UCLA Working Papers in Phonetics*, 112–156.

Kent, R. D., & Netsell, R. (1971). Effects of stress contrasts on certain articulatory parameters. *Phonetica*, *24*(1), 23–44.

Kim, S. (2004). *The role of prosodic phrasing in Korean word segmentation* (Unpublished doctoral dissertation). University of California, Los Angeles.

Kim, S., & Cho, T. (2009). The use of phrase-level prosodic information in lexical segmentation: Evidence from word-spotting experiments in Korean. *The Journal of the Acoustical Society of America*, *125*(5), 3373–3386.

Kim, S., & Cho, T. (2013). Prosodic boundary information modulates phonetic categorization. *The Journal of the Acoustical Society of America*, *134*(1), EL19–EL25.

Kim, S., Choi, J., & Cho, T. (2016, July 13-16). *Linguistic contrast enhancement under prosodic strengthening in L1 and L2 speech.* Poster presented at the 15th Conference on Laboratory Phonology, Ithaca, NY, United States.

Kim, S., Kim, J., & Cho, T. (2018). Prosodic-structural modulation of stop voicing contrast

along the VOT continuum in trochaic and iambic words in American English. *Journal of Phonetics*, *71*, 65–80.

Kim, S., Mitterer, H., & Cho, T. (2018). A time course of prosodic modulation in phonological inferencing: The case of Korean post-obstruent tensing. *PloS one*, *13*(8).

Kingston, J., Levy, J., Rysling, A., & Staub, A. (2016). Eye movement evidence for an immediate Ganong effect. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(12), 1969.

Klatt, D. H. (1975). Vowel lengthening is syntactically determined in a connected discourse. *Journal of Phonetics*, *3*(3), 129–140.

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *The Journal of the Acoustical Society of America*, *59*(5), 1208–1221.

Kohler, K. J. (1994). Glottal stops and glottalization in German. *Phonetica*, *51*(1-3), 38–51.

Krakow, R. A., Bell-Berti, F., & Wang, Q. E. (1995). Supralaryngeal declination: evidence from the velum. In *Producing speech: Contemporary issues* (pp. 333–354). AIP Press.

Krivokapić, J. (2012). Prosodic planning in speech production. In *Speech planning and dynamics* (pp. 157–190). Peter Lang.

Krivokapić, J. (2014). Gestural coordination at prosodic boundaries and its role for prosodic structure and speech planning processes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1658), 20130397.

Krivokapić, J., & Byrd, D. (2012). Prosodic boundary strength: An articulatory and perceptual study. *Journal of Phonetics*, *40*(3), 430–442.

Ladd, D. R. (2008). *Intonational Phonology.* Cambridge University Press.

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *The Journal of the Acoustical Society of America*, *29*(1), 98–104.

Lehet, M., & Holt, L. L. (2020). Nevertheless, it persists: Dimension-based statistical learning and normalization of speech impact different levels of perceptual processing. *Cognition*, *202*, 104328.

Lehiste, I. (1970). *Suprasegmentals.* Massachusetts Institute of Technology Press.

Lehiste, I., & Peterson, G. E. (1959). Vowel amplitude and phonemic stress in American

English. *The Journal of the Acoustical Society of America*, *31*(4), 428–435.

Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2018). *emmeans: Estimated Marginal Means, aka Least-Squares Means.* Retrieved from `https://CRAN.R-project.org/package=emmeans`

Liberman, M., & Prince, A. (1977). On stress and linguistic rhythm. *Linguistic Inquiry*, *8*(2), 249–336.

Liljencrants, J., & Lindblom, B. (1972). Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language*, 839–862.

Lindblom, B. (1968). Temporal organization of syllable production. *Speech Transmission Laboratory Quarterly Progress and Status Report*, *9*(2-3), 1–5.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In *Speech Production and Speech Modeling* (pp. 403–439). Springer.

Lisker, L. (1986). "Voicing" in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees. *Language and Speech*, *29*(1), 3–11.

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, *20*(3), 384–422.

Lisker, L., & Abramson, A. S. (1970). The voicing dimension: Some experiments in comparative phonetics. In *Proceedings of the 6th International Congress of Phonetic Sciences* (pp. 563–567).

Löfqvist, A., Baer, T., McGarr, N. S., & Story, R. S. (1989). The cricothyroid muscle in voicing control. *The Journal of the Acoustical Society of America*, *85*(3), 1314–1321.

Lotto, A. J., Kluender, K. R., & Holt, L. L. (1997). Perceptual compensation for coarticulation by Japanese quail (Coturnix coturnix japonica). *The Journal of the Acoustical Society of America*, *102*(2), 1134–1140.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, *19*(1), 1.

Magen, H. S. (1997). The extent of vowel-to-vowel coarticulation in English. *Journal of Phonetics*, *25*(2), 187–205.

Magnuson, J. S., Mirman, D., Luthra, S., Strauss, T., & Harris, H. D. (2018). Interaction

in spoken word recognition models: Feedback helps. *Frontiers in Psychology*, *9*, 369.

Mann, V. A. (1980). Influence of preceding liquid on stop-consonant perception. *Perception & Psychophysics*, *28*(5), 407–412.

Marge, M., Banerjee, S., & Rudnicky, A. I. (2010). Using the Amazon Mechanical Turk for transcription of spoken language. In *Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 5270–5273).

Marslen-Wilson, W., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition*, *8*(1), 1–71.

Maslowski, M., Meyer, A. S., & Bosker, H. R. (2019). How the tracking of habitual rate influences speech perception. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(1), 128.

Maslowski, M., Meyer, A. S., & Bosker, H. R. (2020). Eye-tracking the time course of distal and global speech rate effects. *Journal of Experimental Psychology: Human Perception and Performance*.

Matin, E., Shao, K. C., & Boff, K. R. (1993). Saccadic overhead: Information-processing time with and without saccades. *Perception & Psychophysics*, *53*(4), 372–380.

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

McMurray, B., Cole, J. S., & Munson, C. (2011). Features as an emergent product of computing perceptual cues relative to expectations. In *Where do phonological features come from?* (pp. 197–236). John Benjamins.

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*(2), 219.

McQueen, J. M., Cutler, A., & Norris, D. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science*, *30*(6), 1113–1126.

Melguy, Y. (2018). Strengthening, Weakening and Variability: The Articulatory Correlates of Hypo-and Hyper-articulation in the Production of English Dental Fricatives. *UC Berkeley PhonLab Annual Report*, *14*(1).

Mendelsohn, A. H., & Zhang, Z. (2011). Phonation threshold pressure and onset frequency in a two-layer physical model of the vocal folds. *The Journal of the Acoustical Society of America*, *130*(5), 2961–2968.

Miller, J. L., & Liberman, A. M. (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel. *Perception & Psychophysics*, *25*(6), 457–465.

Mitterer, H., Cho, T., & Kim, S. (2016). How does prosody influence speech categorization? *Journal of Phonetics*, *54*, 68–79.

Mitterer, H., & Ernestus, M. (2008). The link between speech perception and production is phonological and abstract: Evidence from the shadowing task. *Cognition*, *109*(1), 168–173.

Mitterer, H., Kim, S., & Cho, T. (2019). The glottal stop between segmental and suprasegmental processing: The case of Maltese. *Journal of Memory and Language*, *108*, 104034.

Mitterer, H., & Reinisch, E. (2013). No delays in application of perceptual learning in speech recognition: Evidence from eye tracking. *Journal of Memory and Language*, *69*(4), 527–545.

Mo, Y. (2008). Duration and intensity as perceptual cues for naïve listeners' prominence and boundary perception. In *Proceedings of the 4th International Conference on Speech Prosody* (pp. 739–742).

Mo, Y. (2011). *Prosody production and perception with conversational speech* (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign.

Mo, Y., Cole, J., & Hasegawa-Johnson, M. (2009). Prosodic effects on vowel production: evidence from formant structure. In *Proceedings of INTERSPEECH* (pp. 2535–2538).

Mooshammer, C., & Geng, C. (2008). Acoustic and articulatory manifestations of vowel reduction in German. *Journal of the International Phonetic Association*, 117–136.

Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, *9*(5-6), 453–467.

Mücke, D., & Grice, M. (2014). The effect of focus marking on supralaryngeal articulation–is

it mediated by accentuation? *Journal of Phonetics*, *44*, 47–61.

Nadeu, M. (2014). Stress-and speech rate-induced vowel quality variation in Catalan and Spanish. *Journal of Phonetics*, *46*, 1–22.

Nakai, S., Kunnari, S., Turk, A., Suomi, K., & Ylitalo, R. (2009). Utterance-final lengthening and quantity in Northern Finnish. *Journal of Phonetics*, *37*(1), 29–45.

Nakai, S., & Turk, A. E. (2011). Separability of prosodic phrase boundary and phonemic information. *The Journal of the Acoustical Society of America*, *129*(2), 966–976.

Nearey, T. M. (1997). Speech perception as pattern recognition. *The Journal of the Acoustical Society of America*, *101*(6), 3241–3254.

Nespor, M., & Vogel, I. (2007). *Prosodic Phonology: With a New Foreword*. Walter de Gruyter.

Newman, R. S., & Sawusch, J. R. (1996). Perceptual normalization for speaking rate: Effects of temporal distance. *Perception & Psychophysics*, *58*(4), 540–560.

Newman, R. S., & Sawusch, J. R. (2009). Perceptual normalization for speaking rate III: Effects of the rate of one voice on perception of another. *Journal of Phonetics*, *37*(1), 46–65.

Newman, R. S., Sawusch, J. R., & Luce, P. A. (1997). Lexical neighborhood effects in phonetic processing. *Journal of Experimental Psychology: Human Perception and Performance*, *23*(3), 873.

Nixon, J. S., van Rij, J., Mok, P., Baayen, R. H., & Chen, Y. (2016). The temporal dynamics of perceptual uncertainty: eye movement evidence from Cantonese segment and tone perception. *Journal of Memory and Language*, *90*, 103–125.

Nooteboom, S. G., & Doodeman, G. J. (1980). Production and perception of vowel length in spoken sentences. *The Journal of the Acoustical Society of America*, *67*(1), 276–287.

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition*, *52*(3), 189–234.

Norris, D. (1999). The merge model: Speech perception is bottom-up. *The Journal of the Acoustical Society of America*, *106*(4), 2295–2295.

Norris, D., & McQueen, J. M. (2008). Shortlist B: a Bayesian model of continuous speech

recognition. *Psychological Review*, *115*(2), 357.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences*, *23*, 299–325.

Norris, D., McQueen, J. M., & Cutler, A. (2016). Prediction, Bayesian inference and feedback in speech recognition. *Language, Cognition and Neuroscience*, *31*(1), 4–18.

Norris, D., McQueen, J. M., & Cutler, A. (2018). Commentary on "Interaction in Spoken Word Recognition Models". *Frontiers in Psychology*, *9*, 1568.

Öhman, S. E. (1966). Coarticulation in VCV utterances: Spectrographic measurements. *The Journal of the Acoustical Society of America*, *39*(1), 151–168.

Peterson, G. E., & Barney, H. L. (1952). Control Methods Used in a Study of the Vowels. *The Journal of the Acoustical Society of America*, *24*(2), 175–184.

Peterson, G. E., & Lehiste, I. (1960). Duration of syllable nuclei in English. *The Journal of the Acoustical Society of America*, *32*(6), 693–703.

Pierrehumbert, J. B. (1980). *The phonology and phonetics of English intonation* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.

Pierrehumbert, J. B., & Frisch, S. (1997). Synthesizing allophonic glottalization. In *Progress in Speech Synthesis* (pp. 9–26). Springer.

Pierrehumbert, J. B., & Talkin, D. (1992). Lenition of /h/ and glottal stop. In G. J. Docherty & D. R. Ladd (Eds.), *Gesture, Segment, Prosody* (p. 90-127). Cambridge University Press.

Plantinga, J., & Trainor, L. J. (2005). Memory for melody: Infants use a relative pitch code. *Cognition*, *98*(1), 1–11.

Politzer-Ahles, S., & Piccinini, P. (2018). On visualizing phonetic data from repeated measures experiments with multiple random effects. *Journal of Phonetics*, *70*, 56–69.

Price, P. J., Ostendorf, M., Shattuck-Hufnagel, S., & Fong, C. (1991). The use of prosody in syntactic disambiguation. *the Journal of the Acoustical Society of America*, *90*(6), 2956–2970.

R Core Team. (2020). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Raphael, L. J. (1972). Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English. *The Journal of the Acoustical Society of America*, *51*(4B), 1296–1303.

Recasens, D. (1984). Vowel-to-vowel coarticulation in Catalan VCV sequences. *The Journal of the Acoustical Society of America*, *76*(6), 1624–1635.

Reinisch, E. (2016). Speaker-specific processing and local context information: The case of speaking rate. *Applied Psycholinguistics*, *37*(6), 1397–1415.

Reinisch, E., & Sjerps, M. J. (2013). The uptake of spectral and temporal cues in vowel perception is rapidly influenced by context. *Journal of Phonetics*, *41*(2), 101–116.

Repp, B. H. (1983). Bidirectional contrast effects in the perception of VC-CV sequences. *Perception & Psychophysics*, *33*(2), 147–155.

Repp, B. H. (1997). Spectral envelope and context effects in the tritone paradox. *Perception*, *26*(5), 645–665.

Rietveld, A., & Koopmans-van Beinum, F. J. (1987). Vowel reduction and stress. *Speech communication*, *6*(3), 217–229.

Roessig, S., Mücke, D., & Pagel, L. (2019). Dimensions of prosodic prominence in an attractor model. In *Proceedings of INTERSPEECH* (pp. 2533–2537).

Rysling, A. (2017). *Preferential early attribution in segmental parsing* (Unpublished doctoral dissertation). University of Massachusetts Amherst.

Rysling, A., Bishop, J., Clifton, C., & Yacovone, A. (2020). Preceding syllables are necessary for the accent advantage effect. *The Journal of the Acoustical Society of America*, *148*(63), EL285–EL288.

Rysling, A., Jesse, A., & Kingston, J. (2019). Regressive spectral assimilation bias in speech perception. *Attention, Perception, & Psychophysics*, *81*(4), 1127–1146.

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*(1), 51–89.

Salverda, A. P., Dahan, D., Tanenhaus, M. K., Crosswhite, K., Masharov, M., & McDonough, J. (2007). Effects of prosodically modulated sub-phonetic variation on lexical competition. *Cognition*, *105*(2), 466–476.

Schafer, A. J. (1997). *Prosodic parsing: The role of prosody in sentence comprehension* (Unpublished doctoral dissertation). University of Massachusetts Amherst.

Schellenberg, E. G., & Trehub, S. E. (2003). Good pitch memory is widespread. *Psychological Science*, *14*(3), 262–266.

Schwarzschild, R. (1999). GIVENness, AvoidF and other constraints on the placement of accent. *Natural Language Semantics*, *7*(2), 141–177.

Schweitzer, K., Walsh, M., Calhoun, S., Schütze, H., Möbius, B., Schweitzer, A., & Dogil, G. (2015). Exploring the relationship between intonation and the lexicon: Evidence for lexicalised storage of intonation. *Speech Communication*, *66*, 65–81.

Selkirk, E. (1995). Sentence prosody: Intonation, stress, and phrasing. *The Handbook of Phonological Theory*, *1*, 550–569.

Seo, J., Kim, S., Kubozono, H., & Cho, T. (2019). Preboundary lengthening in Japanese: To what extent do lexical pitch accent and moraic structure matter? *The Journal of the Acoustical Society of America*, *146*(3), 1817–1823.

Seyfarth, S., & Garellek, M. (2018). Plosive voicing acoustics and voice quality in Yerevan Armenian. *Journal of Phonetics*, *71*, 425–450.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, *270*(5234), 303–304.

Shattuck-Hufnagel, S. (2000). Phrase-level phonology in speech production planning: Evidence for the role of prosodic structure. In *Prosody: Theory and experiment* (pp. 201–229). Springer.

Shepherd, M. A. (2008). The scope and effects of preboundary prosodic lengthening in Japanese. *USC Working Papers in Linguistics*, *4*, 1–14.

Shields, J. L., McHugh, A., & Martin, J. G. (1974). Reaction time to phoneme targets as a function of rhythmic cues in continuous speech. *Journal of Experimental Psychology*, *102*(2), 250.

Shultz, A. A., Francis, A. L., & Llanos, F. (2012). Differential cue weighting in perception and production of consonant voicing. *The Journal of the Acoustical Society of America*, *132*(2), EL95–EL101.

Silbert, N., & de Jong, K. (2008). Focus, prosodic context, and phonological feature specification: Patterns of variation in fricative production. *The Journal of the Acoustical Society of America*, *123*(5), 2769–2779.

Silverman, K., & Pierrehumbert, J. (1990). The timing of prenuclear high accents in English. In M. E. Beckman & J. Kingston (Eds.), *Papers in Laboratory Phonology* (pp. 72–106).

Sjerps, M. J., Mitterer, H., & McQueen, J. M. (2011). Constraints on the processes responsible for the extrinsic normalization of vowels. *Attention, Perception, & Psychophysics*, *73*(4), 1195–1215.

Slifka, J. (2006). Some physiological correlates to regular and irregular phonation at the end of an utterance. *Journal of Voice*, *20*(2), 171–186.

Slote, J., & Strand, J. F. (2016). Conducting spoken word recognition research online: Validation and a new timing method. *Behavior Research Methods*, *48*(2), 553–566.

Sóskuthy, M. (2017). Generalised additive mixed models for dynamic analysis in linguistics: a practical introduction. *arXiv preprint arXiv:1703.05339*.

Speer, S. R., Kjelgaard, M. M., & Dobroth, K. M. (1996). The influence of prosodic structure on the resolution of temporary syntactic closure ambiguities. *Journal of Psycholinguistic Research*, *25*(2), 249–271.

Speer, S. R., Warren, P., & Schafer, A. (2003). Intonation and sentence processing. In *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 95–105).

Sprouse, J. (2011). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods*, *43*(1), 155–167.

Steffman, J. (2019a). Intonational structure mediates speech rate normalization in the perception of segmental categories. *Journal of Phonetics*, *74*, 114–129.

Steffman, J. (2019b). Phrase-final lengthening modulates listeners' perception of vowel duration as a cue to coda stop voicing. *The Journal of the Acoustical Society of America*, *145*(6), EL560–EL566.

Steffman, J., & Jun, S.-A. (2019). Perceptual integration of pitch and duration: Prosodic and psychoacoustic influences in speech perception. *The Journal of the Acoustical Society of America*, *146*(3), EL251–EL257.

Steffman, J., & Katsuda, H. (2020). Intonational structure influences perception of contrastive vowel length: The case of phrase-final lengthening in Tokyo Japanese. *Language and Speech*, 0023830920971842.

Steriade, D. (1997). Phonetics in phonology: the case of laryngeal neutralization. *UCLA Working Papers in Linguistics*, 25–146.

Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *The Journal of the Acoustical Society of America*, *111*(4), 1872–1891.

Stilp, C. (2018). Short-term, not long-term, average spectra of preceding sentences bias consonant categorization. *The Journal of the Acoustical Society of America*, *144*(3), 1797–1797.

Stilp, C. (2020). Acoustic context effects in speech perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, *11*(1), e1517.

Stilp, C., Alexander, J. M., Kiefte, M., & Kluender, K. R. (2010). Auditory color constancy: Calibration to reliable spectral properties across nonspeech context and targets. *Attention, Perception, & Psychophysics*, *72*(2), 470–480.

Swerts, M., & Krahmer, E. (2008). Facial expression and prosodic prominence: Effects of modality and facial area. *Journal of Phonetics*, *36*(2), 219–238.

Swinney, D. A. (1979). Lexical access during sentence comprehension: (Re)consideration of context effects. *Journal of Verbal Learning and Verbal Behavior*, *18*(6), 645–659.

Tanenhaus, M. K., Leiman, J. M., & Seidenberg, M. S. (1979). Evidence for multiple stages in the processing of ambiguous words in syntactic contexts. *Journal of Verbal Learning and Verbal Behavior*, *18*(4), 427–440.

Tehrani, H. (2020). *Appsobabble: Online applications platform.* Retrieved from `https://www.appsobabble.com`

Terken, J., & Hermes, D. (2000). The perception of prosodic prominence. In *Prosody: Theory and Experiment* (pp. 89–127). Springer.

't Hart, J., Collier, R., & Cohen, A. (1990). *A Perceptual Study of Intonation: An Experimental-Phonetic Approach to Speech Melody.* Cambridge University Press.

Toscano, J. C., & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science*, *34*(3), 434–464.

Toscano, J. C., & McMurray, B. (2012). Cue-integration and context effects in speech: Evidence against speaking-rate normalization. *Attention, Perception, & Psychophysics*, *74*(6), 1284–1301.

Toscano, J. C., & McMurray, B. (2015). The time-course of speaking rate compensation: Effects of sentential rate and vowel length on voicing judgments. *Language, Cognition and Neuroscience*, *30*(5), 529–543.

Traunmüller, H. (1990). Analytical expressions for the tonotopic sensory scale. *The Journal of the Acoustical Society of America*, *88*(1), 97–100.

Truckenbrodt, H. (1995). *Phonological phrases–their relation to syntax, focus, and prominance* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.

Turk, A. E., & Shattuck-Hufnagel, S. (2007). Multiple targets of phrase-final lengthening in American English words. *Journal of Phonetics*, *35*(4), 445–472.

Umeda, N. (1975). Vowel duration in American English. *The Journal of the Acoustical Society of America*, *58*(2), 434–445.

van Dommelen, W. A. (1993). Does dynamic F0 increase perceived duration? New light on an old issue. *Journal of Phonetics*, *21*(4), 367–386.

van Rij, J., Wieling, M., Baayen, R., & van Rijn, H. (2016). *itsadug: Interpreting time series and autocorrelated data using GAMMs [R package].*

van Summers, W. (1987). Effects of stress and final-consonant voicing on vowel production: Articulatory and acoustic analyses. *The Journal of the Acoustical Society of America*, *82*(3), 847–863.

Vasishth, S., Nicenboim, B., Beckman, M. E., Li, F., & Kong, E. J. (2018). Bayesian data analysis in the phonetic sciences: A tutorial introduction. *Journal of Phonetics*, *71*, 147–161.

Vitevitch, M. S., & Luce, P. A. (1998). When words compete: Levels of processing in perception of spoken words. *Psychological science*, *9*(4), 325–329.

Vitevitch, M. S., Luce, P. A., Pisoni, D. B., & Auer, E. T. (1999). Phonotactics, neighborhood activation, and lexical access for spoken words. *Brain and Language*, *68*(1-2), 306–311.

Wade, T., & Holt, L. L. (2005). Perceptual effects of preceding nonspeech rate on temporal properties of speech categories. *Perception & Psychophysics*, *67*(6), 939–950.

Wagner, M., & Crivellaro, S. (2010). Relative prosodic boundary strength and prior bias in disambiguation. In *Proceedings of the 5th International Conference on Speech Prosody.*

Wagner, P., Origlia, A., Avezani, C., Christodoulides, G., Cutugno, F., d'Imperio, M., ... others (2015). Different parts of the same elephant: A roadmap to disentangle and connect different perspectives on prosodic prominence. In *Proceedings of the 18th International Congress of Phonetic Sciences.*

Watson, D. G., Arnold, J. E., & Tanenhaus, M. K. (2008). Tic tac toe: Effects of predictability and importance on acoustic prominence in language production. *Cognition*, *106*(3), 1548–1557.

Whalen, D. H., Abramson, A. S., Lisker, L., & Mody, M. (1993). F0 gives voicing information even with unambiguous voice onset times. *The Journal of the Acoustical Society of America*, *93*(4), 2152–2159.

White, L., Benavides-Varela, S., & Mády, K. (2020). Are initial-consonant lengthening and final-vowel lengthening both universal word segmentation cues? *Journal of Phonetics*, *81*, 100982.

Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. J. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *The Journal of the Acoustical Society of America*, *91*(3), 1707–1717.

Winn, M. (2016). *Vowel formant continua from modified natural speech (Praat script).* Retrieved from `http://www.mattwinn.com/praat/Make_Formant_Continuum_v38.txt` (Version 38)

Wood, S. N. (2017). *Generalized Additive Models: an Introduction with R.* Chapman and Hall/CRC.

Xu, L., Thompson, C. S., & Pfingst, B. E. (2005). Relative contributions of spectral

and temporal cues for phoneme recognition. *The Journal of the Acoustical Society of America*, *117*(5), 3255–3267.

Xu, L., & Zheng, Y. (2007). Spectral and temporal cues for phoneme recognition in noise. *The Journal of the Acoustical Society of America*, *122*(3), 1758–1764.

Yao, Y., Tilsen, S., Sprouse, R. L., & Johnson, K. (2010). Automated Measurement of Vowel Formants in the Buckeye Corpus. *UC Berkeley PhonLab Annual Report*, *6*(6).

Yu, A., Lee, H., & Lee, J. (2014). Variability in perceived duration: pitch dynamics and vowel quality. In *Proceedings of the 4th International Symposium on Tonal Aspects of Languages* (pp. 41–44).

Zahner, K., Kutscheid, S., & Braun, B. (2019). Alignment of f0 peak in different pitch accent types affects perception of metrical stress. *Journal of Phonetics*, *74*, 75–95.

Zhang, Z. (2011). Restraining mechanisms in regulating glottal closure during phonation. *The Journal of the Acoustical Society of America*, *130*(6), 4010–4019.