Prosodic prominence effects in the processing of spectral cues

Jeremy Steffman[a]

[a]UCLA

ABSTRACT

Two experiments test how phrasal prominence influences listeners' perception of vowel contrasts and how prominence information and vowel formant cues are integrated in processing. Experiment 1 finds that listeners incorporate phrasal prominence in their perception of vowels, in line with how spectral structure is modulated by prominence in speech. Experiment 2 explores how prominence information is integrated with formant cues in a visual world eyetracking task. Prominence shows an overall later influence in processing in line with current models of prosodic and segmental integration. However, listeners' perception of formants was also impacted more subtly by prominence immediately in processing such that prominence information directly shapes how formant cues are perceived. Results are discussed in terms of their implications for models of prosodic effects in segmental perception and possible differences between prosodic prominence and prosodic boundaries in this regard.

KEYWORDS

Speech perception; prosody; prominence; vowel perception

1. Introduction

Two key parts of understanding spoken language for listeners are (1) perceiving contrastive segmental categories which convey lexical distinctions in their language, and (2) perceiving prosodic categories which convey various other pieces of information, such as phrasal grouping, prominence relations, information structure, and so on (Cho, McQueen, & Cox, 2007; Cutler, Dahan, & Van Donselaar, 1997; Mitterer, Kim, & Cho, 2019). Traditionally, models of spoken word recognition focus on the former of these two tasks, though recent studies have suggested that phrasal prosodic boundaries play

a role in segmental processing (Kim, Mitterer, & Cho, 2018; Mitterer et al., 2019), suggesting prosodic and segmental information is integrated in this regard. However, it is currently unknown to what degree prosodic prominence might play a similar role, and how prominence and segmental information are integrated in processing. The present study accordingly tests how phrasal prosodic prominence mediates listeners' perception of vowel contrasts, and how prominence information is integrated in segmental perception online.

Phrasal prosody above the word is typically studied in the context of syntactic and post-lexical processing (Cutler et al., 1997; Price, Ostendorf, Shattuck-Hufnagel, & Fong, 1991). Accordingly, relatively little is known about the effects of phrasal prosody on the perception of speech sounds. Why should context related to the phrase as a whole (i.e. grouping above the word, phrasal prominence, intonational tunes and structure) matter for perception of segmental contrasts? A large body of phonetic research shows that prosodic prominence and boundaries influence the phonetic realization of segments, conceptualized as the phonetic encoding of prosodic structure (Keating, 2006). The general idea is that phrasal prosodic organization fine-tunes both timing and amplitude patterns in segmental articulations (see e.g., Cho 2015, 2016 and Keating 2006), which helps enhance both paradigmatic and syntagmatic contrasts in segmental articulations (Cho, 2005). In this sense, prosody is seen as introducing systematic variation in segmental structure, with acoustic consequences that may be relevant to listeners. This sort of variation might be particularly relevant to listeners for cues that are used to make segmental contrasts.

Consider a well-known example. One general phenomenon in the literature related to prosodic boundaries is *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 the middle of a prosodic domain (Fougeron & Keating, 1997; Keating, Cho, Fougeron, & Hsu, 2004). One acoustic consequence of initial strengthening is manifested in Voice Onset Time (VOT). In aspirated stops, VOT is longer at the beginning of an Intonational Phrase (IP)[1], as compared to being IP-medial (Cho & Keating, 2001, 2009; Jun, 1996).

[1]Here the terminology from the prosodic hierarchy described in Beckman and Pierrehumbert (1986) and

In this sense VOT duration co-varies in a systematic way with prosodic context. VOT is additionally a strong cue to laryngeal contrasts, where, for example longer VOT cues a voiceless stop (in e.g., English) or an aspirated stop (in e.g., Thai). That is, VOT is clearly an important property for conveying contrasts at the segmental level (e.g.,Abramson 1976; Abramson and Whalen 2017), though it simultaneously patterns in a systemic way as a function of phrasal prosody, i.e. it "encodes" phrasal prosodic structure (Keating, 2006).

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 can cue segmental contrasts, listeners would benefit from reconciling a cue with the prosodic context in which it occurs in perceiving segmental categories. 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). In the context of speech perception at the segmental level, prosodically driven variation in a cue that signals a segmental contrast (such as VOT) should lead listeners to account for prosodic context in segmental perception, that is, taking into account how a given cue has been impacted by phrasal prosodic factors. This would represent a reconciliation of a contextual influence with a cue value, similar in spirit to e.g., compensation for coarticulation (Mann, 1980). In the case of VOT, this would entail accounting for the fact that lengthened VOT could be due to the presence of an IP boundary, therefore adjusting categorization of VOT on that basis. One way to study the relationship between segmental and phrasal influences in speech perception is accordingly to manipulate prosodic context in some way, and observe its impact on the perception of a given cue (e.g., VOT).

Kim and Cho (2013) tested this aforementioned pattern in American English, and found that listeners shifted their perception of VOT continuum ranging from /pɑ/ to /bɑ/. When the target was cued as IP initial by preceding changes in duration and pitch in a carrier phrase, overall longer VOT was required for a voiceless percept,

Pierrehumbert (1980) is adopted, in keeping with much of the literature on domain-initial strengthening.

suggesting listeners took IP initial lengthening into account in their perception of the voicing contrast (see Mitterer, Cho, and Kim (2016) and Steffman (2019b) for further discussion). Since Kim and Cho (2013), various other studies have corroborated this idea, showing that listeners seemingly reference prosodic structure in their perception of various segmental cues, in line with how those cues vary based on prosody in speech production. This has been shown both for domain-initial and domain-final boundary effects, as in phrase-final, or pre-boundary, lengthening (Katsuda & Steffman, 2020; Mitterer et al., 2019; Steffman, 2019b). 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, these findings have focused on the influence of prosodic boundaries.

1.1. How are prosody and segment integrated in perception?

The general findings mentioned above show that listeners evidently integrate prosodic information, in some way, in segmental processing. However this leaves open the question of how both prosodic structure and segmental categories are being processed.

Cho et al. (2007) propose a model, the *Prosody Analyzer*, to explain the influence of prosodic structure in listeners' word segmentation and lexical access. On the basis of data from a cross-modal priming experiment in which phrasal (IP) boundaries 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. The Prosody Analyzer model 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, and is combined with a prosodic representation, which encodes boundary information, thus facilitating segmentation. The role of prosody more generally in this model is to help select between possible word candidates (which are activated on the basis of segmental information), as a function of the context in which they occur.

The model therefore demarcates the role of prosodic and segmental representa-

tions as entering at different stages in the process of word recognition (though note again that this is agnostic as to what listeners use as an acoustic cue for the purpose of specifying each type of representation). 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 general findings showing that word recognition integrates various sources of information in multiple stages (see e.g., Luce and Pisoni 1998 for a general overview). Some of these influences have been explicitly argued to modulate competition at a later stage in word recognition, such as neighborhood density (Newman, Sawusch, & Luce, 1997; Vitevitch & Luce, 1998; Vitevitch, Luce, Pisoni, & Auer, 1999), and semantic context (Cairns & Hsu, 1980; Swinney, 1979), thus there is clear precedent for integration of non-segmental information with segmental information at a later stage. 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 such as in syntactic processing (see discussion in Cho et al. 2007).

Two recent studies (Kim et al., 2018; Mitterer et al., 2019) present time-course evidence in support of this sort of later-stage prosodic analysis. Kim et al. (2018) tested how Korean listeners use tonal prosodic cues in a phonological inferencing task. They tested Korean post-obstruent tensing (POT), whereby lax stops and affricates become tense following another obstruent. To use an example from the paper: /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). The authors tested if listeners, in a visual world eyetracking task, 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 phrasing to determine whether POT may have applied.[2] Kim et al. found that when listeners heard a phonetically tense target (e.g., [p*uri]),

[2]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".

they looked more to an underlyingly lax word (e.g., /puri/), when that word was in an AP-internal context that licensed POT, as compared to when it was not. This evidenced the predicted phonological inferencing effect. The authors further showed that this effect goes away when an AP boundary intervenes between the 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 "[...] prosodic structure is parsed in parallel to the segmental level and is used later for prosodic modulation in lexical access" (p26).

Mitterer et al. (2019) tested how listeners might use prosodic boundaries to modulate perception of a segment 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, Shattuck-Hufnagel, and Ostendorf 1996; Pierrehumbert and Talkin 1992). Mitterer et al. (2019) found that listeners used preceding boundary information in determining if glottalization cued a segmental contrast (in similar fashion to VOT as described above). However these effects disappeared in 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 prosodic factors. 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 an offline 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. This offers further support to the idea that prosodic boundary effects should occur at a later stage in the word recognition process, too late to be observed online with Mitterer et al's materials online. 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, with prosodic information being used at a later stage.

This, taken together with Kim et al. (2018), offers clear support for a later-stage influence of prosodic information in speech processing, supporting the model proposed by Cho et al. (2007). As summarized by Mitterer et al: "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" (p 14).

1.2. 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?

At the phonological level, prominence in a language like American English can be described as docking on metrically strong syllables (Hayes, 1995; Liberman & Prince, 1977; Nespor & Vogel, 2007) , and further determined by information and discourse structure, e.g., contrast, given-ness, etc. (Beckman & Pierrehumbert, 1986; Bolinger, 1958, 1961; Pierrehumbert, 1980). In this sense "prominence" is configurational, and 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 prominent (i.e., accented) syllable within a phrase (e.g., Cole et al. 2019). This prominence is "structural" in the sense that it depends on metrical structure, prosodic

phrasing, and the position of a prominent syllable within a phrase. In other words, prominence structure of this kind is clearly manifested in the prosodic organization of the phrase as a whole, the computation of which might be supposed to be analogous to, or part of the same structure as, that computed for prosodic boundaries. In this regard, we might expect phrasal prominence processing at this more abstract level might play out in similar fashion to the processing of prosodic boundaries.

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 features, and part of speech information (Baumann & Winter, 2018; Bishop, 2017; Calhoun, 2007; Cole et al., 2019). This makes characterizing the concept of prominence in a complete manner a difficult task(see e.g., Baumann and Cangemi 2020; Wagner et al. 2015). Even if we restrict ourselves to acoustic features of the speech signal, defining prominence is complicated. One 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 shape prominence perception in a granular fashion, i.e. beyond the distinction of nuclear-accented/accented/unaccented encoded in models such as Pierrehumbert (1980). For example, as shown in various rapid prosody transcription (RPT) studies, changes in pitch and duration strongly predict listeners' perception of prominence (Cole, Mo, & Hasegawa-Johnson, 2010; Mo, 2008, 2011). Increases in pitch, and duration generate increases in perceived prominence, as indexed by a so-called P-score in an RPT task, including within pitch accent categories (Bishop, Kuo, & Kim, 2020).[3] This shows that listeners rely on more fine-grained detail than e.g., an accented/unaccented dis-

[3]A P-score is the proportion of times naïve listeners perceive a given word as being prominent, which they annotate as they listen to a speech sample, see e.g. Cole and Shattuck-Hufnagel (2016).

tinction when they perceive prominence (see also e.g., Fant and Kruckenberg 1989; Katz and Selkirk 2011; Mücke and Grice 2014 for similar arguments). 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 prosodic context) shaped how listeners perceived vowel duration as a cue to coda 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, not directly related to phonological organization in a phrase, can also shape listeners' perception of segmental cues. Importantly though, even this sort of phonetic prominence is relational, in line with the definition given above. Pitch and duration should be perceived as phonetically prominent, or not, relative to context.[4] We could accordingly conceptualize phonetic prominence perception as entailing the relation of a given linguistic unit to its context, e.g. in contrasting relative differences in duration and pitch (Diehl & Walsh, 1989; Repp, 1997). This sort of prominence perception may not involve reference to an abstract (phonological) prosodic structure, but instead may be related to more general acoustic/phonetic prominence, presenting a possible difference from boundary processing. Accordingly, we might expect to see different patterns of processing as compared to those predicted by prosodic analysis, though this is an entirely open question.

How might listeners integrate phonetic and phonological prominence information with their perception of segmental cues? If phonetic prominence is considered an acoustic contextual effect, drawing on contrast mechanisms in perception, we might expect to see this sort of prominence information integrated rapidly, given that other contrast effects show rapid incorporation and are generally thought to operate early in processing (e.g., Bosker et al. 2017; Lehet and Holt 2020; Reinisch and Sjerps 2013; Wade and Holt 2005). This predicts a different timecourse from that of prosodic analysis. One general model for context effects in the literature is C-CuRE ("computing cues relative

[4]For example, the finding that pitch perception is relative to pitch range and context is well established in the psycho-acoustic literature (Plantinga & Trainor, 2005; Repp, 1997; Schellenberg & Trehub, 2003), and the context-dependence of duration perception is also well-established (Bosker, Reinisch, & Sjerps, 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 this word would be relative to stimuli heard on other trials, i.e. the global context of the experiment (cf. Bigand and Pineau 1997; Jones and McAuley 2005).

to expectations"; Cole, Linebaugh, Munson, and McMurray 2010; McMurray, Cole, and Munson 2011; McMurray and Jongman 2011). This model allows for cues to be represented non-veridically to listeners, whereby they may instead be re-coded based on their deviation from expectations generated from context (e.g., a male or female talker). McMurray and Jongman (2011) show that C-CuRE can retain phonetic detail but also allow for perception cues to be modulated by pre-established categories (in model terms, a cue value is re-coded in terms of its deviation from an expected value, derived from by-category regressions). C-CuRE was also shown to provide the best fit to some speech perception data in comparison to a model created by the authors that made use of many cues without compensation (similar in spirit to e.g., Hawkins 2003; Nearey 1997).

The timecourse of these sorts of effects has been demonstrated to be rapid (Reinisch & Sjerps, 2013; Toscano & McMurray, 2015). For example, Toscano and McMurray (2015) tested how VOT, and the speech rate of the material that preceded it, influenced perception of voicing in American English (given that perception of VOT is influences by speech rate). In a visual world eyetracking task, they found that the influence of speech rate and VOT were *simultaneous*. 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. 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 value via expectations generated from speech rate. More generally then, the influence of such a contextual factor should therefore be more or less simultaneous with the cue that it modulates.

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 that of lexical activation. On the other hand, if prominence directly modulates (or re-codes, follow-

ing C-CuRE) cue values in perception, it should show a relatively early influence in perception, in tandem with segmental cues (analogous to the effect of speech rate and VOT in Toscano and McMurray 2015).

Given the complex and multidimensional nature of prominence, which reflects phrasal (phonological) organization, but also derives from acoustic/phonetic context, 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 context effects. Observing how prosodic prominence is processed will accordingly build our current understanding of prosodic context effects in perception, and listeners' integration of prosodic and segmental structure. More precise time-course predictions are discussed in Section 4.

2. The present study

The goal of the two experiments presented in here is to test how phrase-level prominence mediates listeners' use of segmental information, in particular formant cues to a vowel contrast. Vowel articulations, and their acoustic consequences, are modulated in a systematic way by phrasal prominence (outlined below). 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 phrasal prosodic factors. One basic question addressed here is accordingly if listeners are sensitive to this pattern in perception. If the answer is yes, this would provide a first piece of evidence for the involvement of phrasal prominence in segmental perception. This would further extend past findings, which primarily test durational cues, to test 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, and how they combine, following the two accounts sketched above. This will

be discussed in Section 4.

2.1. The test case: Sonority expansion in vowel articulations

Various previous studies have shown that vowels change in their acoustic structure as a function of word level and phrase level prominence (Cho, 2005; Garellek & Keating, 2011; Mooshammer & Geng, 2008; Nadeu, 2014; Van Summers, 1987). These effects are generally seen as serving the purpose of *syntagmatic and paradigmatic contrast enhancement* (Cho, 2005; de Jong, 1991, 1995; de Jong, Beckman, & Edwards, 1993; Roessig, Mücke, & Pagel, 2019). Syntagmatic enhancement effects are those that help a given vocalic articulation contrast with adjacent segments. Paradigmatic effects are those that help a vowel 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).The 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 syntagmatic contrast enhancement is *sonority expansion* (Cho, 2005; de Jong et al., 1993). In this context, sonority is commonly defined in articulatory terms, following Silverman and Pierrehumbert (1990) as "the overall openness of the vocal tract or the impedance looking forward from the glottis" (p. 75). Accordingly, 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. 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).[5]

[5]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., hyper-articulation, are observed (de Jong, 1991, 1995).

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, in a rapid prosody transcription (RPT) task, observed how changes in formant structure influenced how prominent words sounded to listeners. 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 at the level of the phrase) and a segmental dimension (contrastive vowel categories). This is analogous to the case of VOT described above, 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 shifts listeners' perception of F1 and F2. Experiment 1 accordingly tests if listeners incorporate phrasal prominence in their perception of formants.

3. Experiment 1

3.1. 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 contrast chosen as a test case is 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. It should 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 (Umeda, 1975), a point that will be discussed below. Nevertheless, the distinction between these vowels in terms of F1 and F2 is robust, and it will accordingly be assumed that listeners will use F1 and F2 to distinguish these vowel categories, with higher F1 and lower F2 signaling /æ/.

In Experiment 1, listeners' task was to categorize a sound drawn from a continuum as "ebb" /ɛ/ or "ab" /æ/. 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/, produced 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. These two conditions, referred to as the NPA and post-focus condition, were created by cross-splicing, and PSOLA method synthesis, 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

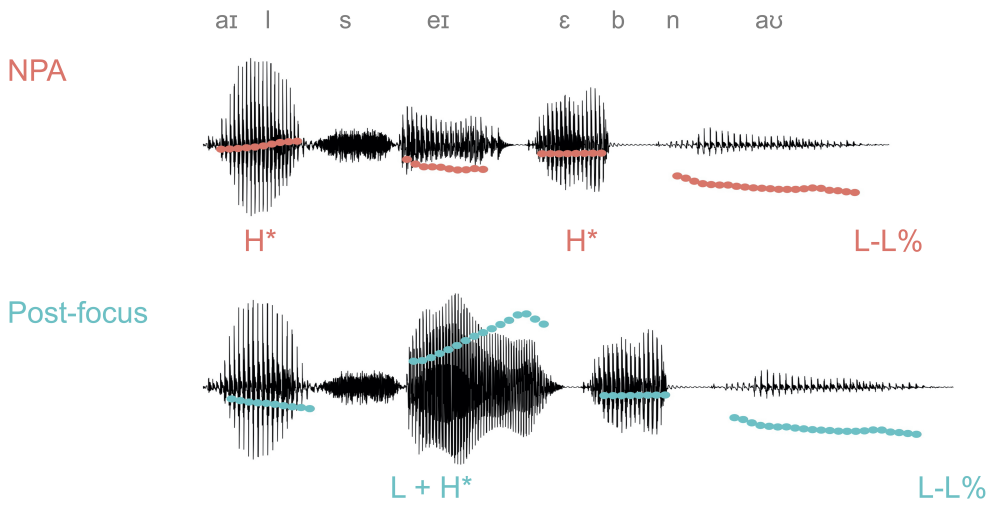

Figure 1. Waveforms of the Experiment 1 stimuli, overlaid with pitch tracks in both conditions. The "nuclear pitch accent" (NPA) condition, in which the target is prominent is shown at the top. The post-focus condition, in which the target is non-prominent, is shown at the bottom. A segmental transcription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

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 (i.e., contextual prominence).

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 following 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

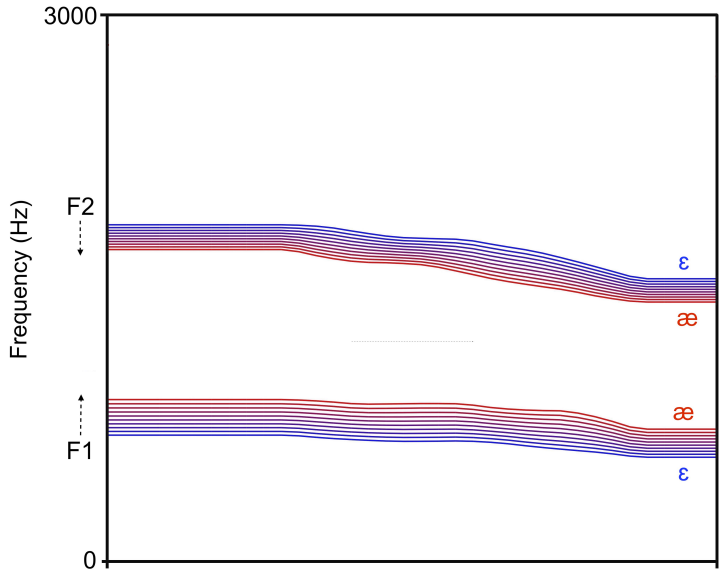

Figure 2. Formant tracks of the Experiment 1 continuum, with F1 and F2 shown. The outermost blue lines represent the formant values for the /ɛ/ endpoint of the continuum (mean F1 = 680 Hz, mean F2 = 1724 Hz). The innermost red lines represent the formant values for the /æ/ endpoint of the continuum (mean F1 = 838 Hz, mean F2 = 1596 Hz). The x axis is time, and is approximately 170 ms in duration.

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 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 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 (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. Each continuum step was then cross-spliced into both NPA and post-focus frames, creating 20 unique stimuli in total (10 continuum steps × 2 frames).

3.2. 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, i.e., jaw lowering and backing of lingual articulation. If listeners attribute these formant changes to prominence (i.e. being driven prosodically instead of segmentally), 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 1, a longer vowel /eɪ/ precedes the target in the post-focus condition, as compared to the NPA condition. Following standard durational contrast effects (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 cue by listeners. 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, following Mitterer et al. (2016) and Steffman (2019a, 2019b).

3.3. Participants and procedure

30 participants were recruited for Experiment 1. All were self-reported native 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 platform that was used to control stimulus presentation in Experiment 1, and all offline categorization experiments (that is, experiments that did not use an eyetracker) 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 a 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 on the computer monitor, 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). Trials were self-paced and there was not any time pressure to provide a response. Halfway through the test trials, participants were prompted to take a short self-paced break. The experiment took approximately 15-20 minutes to complete.

3.4. Results and discussion

Statistical assessment of the categorization responses in Experiment 1 was carried out using a Bayesian logistic mixed effects regression model implemented in the *brms* package in R (Bürkner et al., 2017). The default prior distribution, an improper uniform distribution over real numbers, was used. The reader is referred to Bürkner et al. (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. The output of the Bayesian analysis 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 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

Table 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	

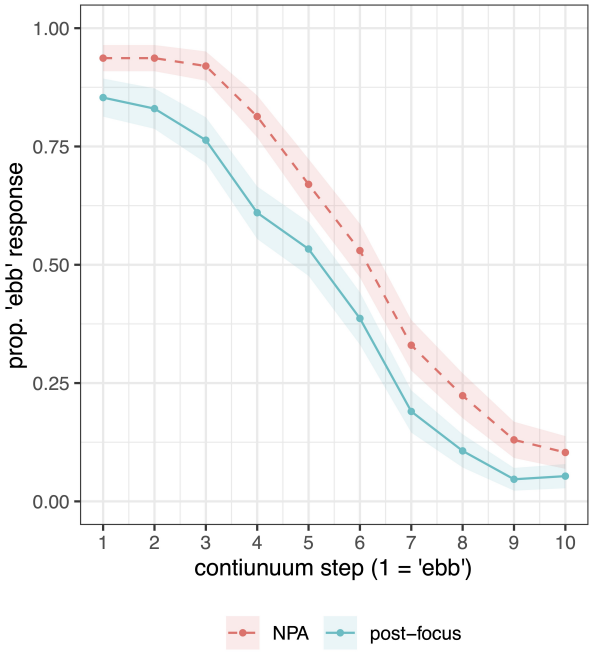

Figure 3. Categorization responses in Experiment 1, with the proportion of "ebb" responses plotted on the y axis, split by prominence condition and continuum step, where step 1 is the /ɛ/ endpoint of the continuum. Shading around each point shows 95% CI.

and random slopes for both fixed effects and their interaction. The model output is shown in Table 1. Categorization responses are plotted in Figure 3.

As shown in Table 1, continuum step impacted categorization such that as continuum 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 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 3, where the categorization function is shifted across prominence conditions.

This findings supports the predictions laid out above: contextual prominence, as cued by phrasal organization and intonation, shifted listeners' perception of the target sound such that they more readily categorized a prominent target as "ebb". This findings 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 general auditory durational contrast (cf. Mitterer et al. 2016) as discussed above in section 1.2.

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, exploring the processing questions outlined above.

4. 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, 1.2 are considered: (1) later-stage modulation of lexical competition by prosodic structure (Kim et al., 2018; Mitterer et al., 2019), and (2) immediate compensation, re-coding a cue value 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. 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 and Reinisch 2013; Toscano and 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 contrast. Formants in a vowel can further be characterized as an *intrinsic* cue, that is, 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 lexical access model sketched above, they should contribute to the early stages of lexical activation, 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.

Experiment 1 also showed that phrasal prominence shaped listeners' perception of the /ɛ/-/æ/ contrast. As described in Section 3.1, the target word was acoustically identical across conditions in this experiment, such that the prominence-lending nature of the 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 stimulus context preceded the target sound in time.

We can now re-frame the two processing 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 et al., 2018; Mitterer et al., 2019) make a clear prediction. Phrasal prosody should exert a later-stage influence, being integrated in the process of lexical competition. This, in relation to vowel-intrinsic formant cues, should occur at a later point in time. The use of prominence information should be *asynchronous* with the use of formant cues.

On the other hand, if prominence immediately modulates perception of the target sound via expectations generated by preceding material in the carrier phrase, and com-

pensatory 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 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 lexical activation in its early stages.

These timecourse predictions are summarized in Table 2.

Table 2. Timecourse predictions for the usage of formants and context

Mechanism	Order of cue usage	Explanation
prosodic analysis	formants cues before prosodic context	formants activate lexical hypotheses and prosodic context subsequently modulates lexical competition
phonetic context	simultaneous	preceding prosodic context modulates expectations about formant values, and is incorporated as soon as formant information becomes available

Consider another difference in processing 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 lexical activation *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. Both this prediction, and the general timing predictions outlined in Table 2 are tested in Experiment 2.

4.1. 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 obtained. 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 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 frames).

4.2. Participants and procedure

36 participants were recruited for Experiment 2 from the same population as Experiment 1. All were self-reported native English speakers, with normal hearing and 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 eye-tracking task in which participants 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 each experiment, participants' gaze was calibrated with a 5-point calibration procedure.

Stimuli were presented binaurally via a PELTOR™3M™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 the monitor. The target words themselves were displayed in 60pt black Arial font and were centered in the left, and right half of the computer screen. 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 target words, to ensure that looks in the vicinity of the target words were also recorded, following e.g., Chong and Garellek (2018).

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.

There were a total of four practice trials in the experiment as in Experiment 1, with each continuum endpoint (from the truncated continuum used in Experiment 2) 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.

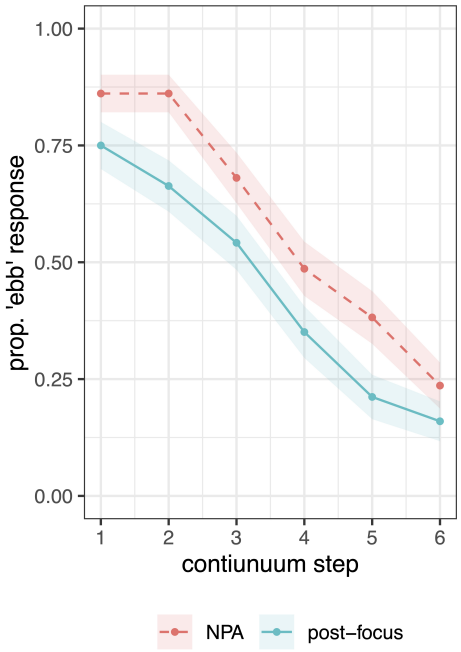

Figure 4. Categorization (click) responses in Experiment 2, showing the proportion of "ebb" responses on the y axis, split by prominence condition and continuum step. Note that steps 1-6 in Experiment 2 correspond to steps 3-8 in Experiment 1, as described in the text. Shading around each point shows 95% CI.

4.3. Results and discussion

4.3.1. Click responses

Listeners' click responses 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, replicating Experiment 1. The model output is shown in Table 3, and categorization responses are plotted in Figure 4. As expected, increasing the continuum step decreased clicks on "ebb", as in Experiment 1 (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.

Table 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	

4.3.2. Eye movement data

Results for eye movement data are shown in Figure 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 5 is simply the proportion 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 of influence 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.[6] With this measure, a negative preference for "ebb" accordingly corresponds to a preference for "ab".

As can be seen in both panels of Figure 5, this preference measure is zero at the beginning of the visible time window, indicating that listeners do not have an

[6]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 expected (cf. Kingston et al. 2016).

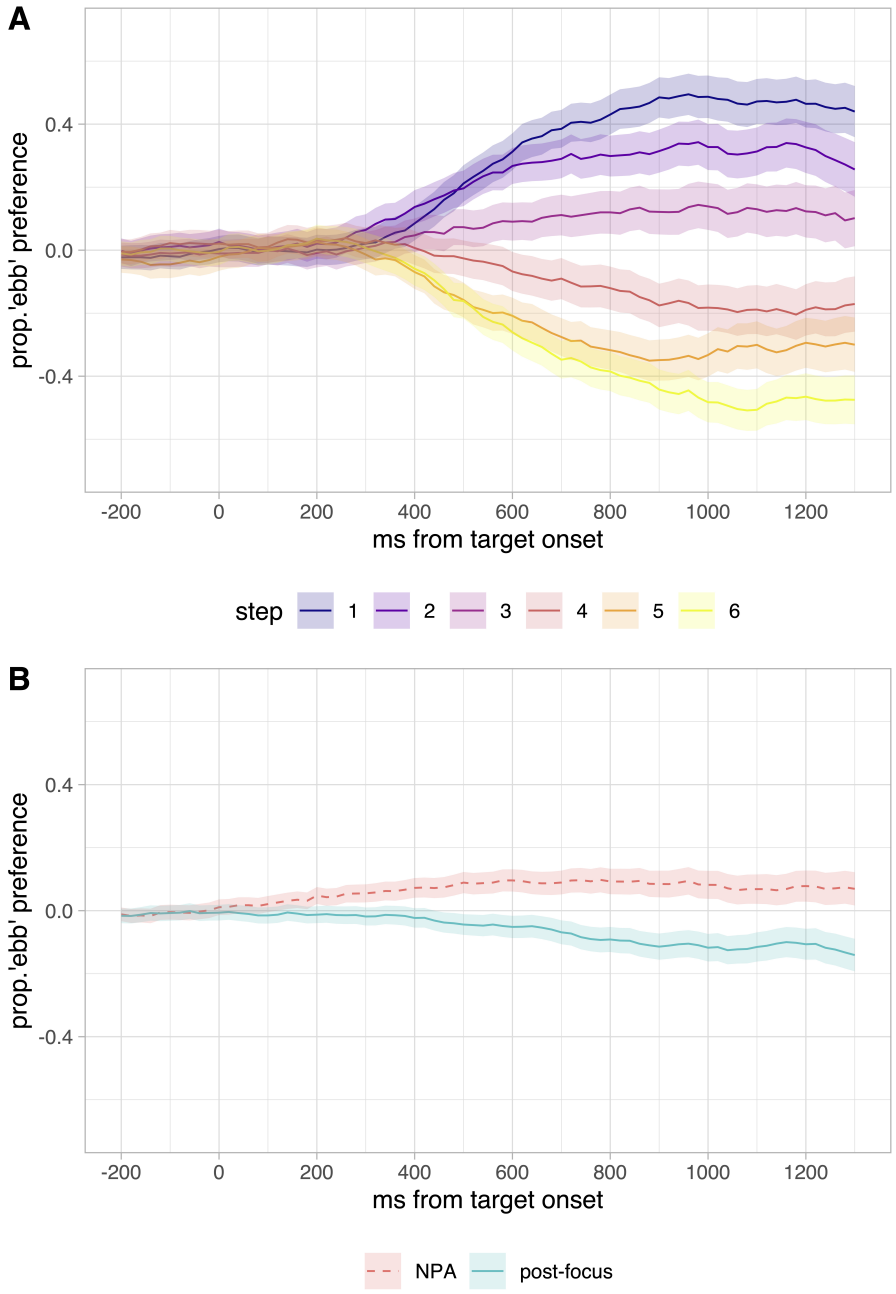

Figure 5. Eye movement data for the effect of continuum step (panel A), and prominence manipulation (panel B), in Experiment 2. The x axis shows time ranging from -200 to 1200 milliseconds from the onset of the target sound. The y axis shows the proportion of looks to "ebb" minus the proportion of looks to "ab" (see text). Confidence regions around each line represent 95% confidence intervals, calculated from the raw data.

immediate preference for either target at the onset of the stimulus. 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 time course results. In the top panel of Figure 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 stronger 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 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 in the analysis window. Notably, divergence on the basis of prominence appears to start early: looks begin to pull apart robustly around 250 ms, though the effect does not appear to reach a stable maximum until much later.

The time course of both of these effects was assessed by a General Additive Mixed Model (GAMM). GAMMs have been applied in various analyses of visual world 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. The reader is referred to Nixon et al. (2016) and Zahner, Kutscheid, and Braun (2019) for discussion of advantages of GAMMs in modeling visual world data, and to Sóskuthy (2017) for a more general overview of GAMMs.

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 two item 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).

The model was fit 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 condition, over time, allowing us to assess how listeners' preferences for a target develop 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 (Sóskuthy, 2017). 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.[7] The model output with both parametric and smooth terms is shown in Table 4. 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).[8] 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 down-sampled to 50 Hz (20ms bins), allowing for a fairly granular timecourse, 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

[7]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.

[8]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.

Table 4. Model output for the GAMM used in Experiment 2, with parametric terms shown above and smooth terms shown below. Notes that p-values for smooth terms indicate whether a smooth is different from a linear predictor.

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	0.05
Smooth terms	edf	ref df	F-value	p-value
te(time, continuum; NPA cond.)	17.09	19.71	38.27	< 0.001
te(time, continuum; post-focus cond.)	8.99	9.52	66.32	< 0.001
s(time, participant; NPA cond.)	228.11	323.00	3.91	< 0.001
s(time, participant; post-focus cond.)	231.32	323.00	4.97	< 0.001

was employed after inspection of the baseline model, as it greatly reduced autocorrelation as compared to a non-AR1 variant (Nixon et al., 2016; Sóskuthy, 2017).[9]

The parametric terms in the model, which represent the overall effect in the analysis window, indicate that both prominence condition and continuum step had an effect on listeners' preference for each target word. As can be seen in Figure 5, increasing the 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, but 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 values along the continuum, differences between smooths of interest were inspected (Sóskuthy, 2017; Zahner et al., 2019). 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

[9]AR1 models assume that neighboring observations in 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.

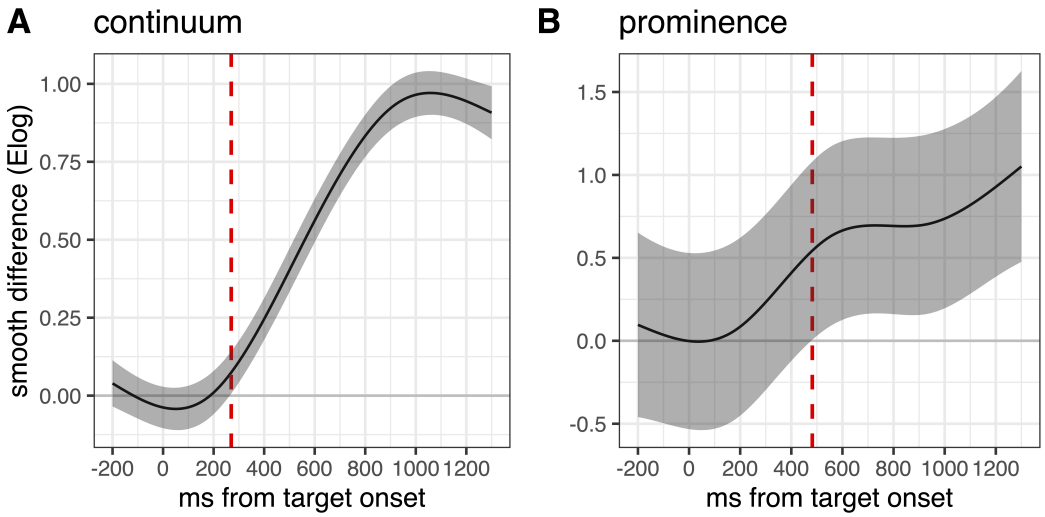

Figure 6. Difference smooths for the effect of continuum step (panel A), and the prominence manipulation (panel B) in Experiment 2. Smooths are surrounded by 95% CI, and the red dashed vertical line indexes when in time CIs exclude zero, i.e. when a difference between the smooths becomes reliable.

expect to see the earliest effect of prominence, a conservative test given the possibility that the effect will be later in processing. This is shown in panel B of Figure 6. The effect of continuum step was assessed by measuring the time of divergence for the two continuum steps which spanned the most ambiguous region in the continuum (steps 3 and 4). Given that each step has its own trajectory, pairwise differences between steps may not necessarily be the same. As can be observed in Figure 5, this area shows the most robust pairwise difference, and the processing of acoustic information at these steps make a relevant comparison to the effect of prominence, which was calculated in the middle of the continuum.[10] This thus represents how quickly formant information is used to distinguish vowels, though notably the estimates for the rest of the pairwise differences between steps only differed by 10-20 ms.[11] This effect is shown in panel A of Figure 6.

As can be seen in Figure 6, the points at which formants along the continuum, and phrasal prominence, impact listeners' target preference reliably are asynchronous. The effect of the continuum causes looks to diverge at a point in time that precedes

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

[11]The divergence estimate shown below is in given for the post-focus condition (as collapsing across conditions is not possible); the effect in the NPA condition showed a similar timecourse.

the divergence of smooths based on phrasal prominence. 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 6 when CI for the model estimate do not include zero, indexed by a dashed vertical line.[12] Another possible way of operationalizing the effect 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 expected to be early, but instead 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: F1 and F2 values should lead to early and immediate activation of lexical candidates, and prosody should mediate candidate 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 previously discussed effect of rate and VOT found in Toscano and McMurray

[12]This timing asynchrony was also seen in a more traditional moving window analysis, not included here. In that analysis time was binned into 100ms windows and a linear mixed effects regression on logit-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 et al. (2018).

2015). 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. 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 (prosodic analysis in the present results) and an intrinsic cue, which is used rapidly.

Though these results would therefore suggest the effect of prominence is relatively late in processing, we can see that the difference between smooths begins to increase well before this point in time. This is also apparent in the raw data, shown in Figure 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. To explore the possibility of an earlier influence of prominence shaping listeners' use of formant cues, as discussed in the predictions above, the non-linear interaction between continuum step, time, and condition (which would evidence an asymmetrical influence of continuum step across conditions) was inspected. 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.

To assess this interaction between continuum step, prominence condition, and time, three dimensional topographic surface plots were used. 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 is represented on a gradient color scale on the z axis. In Figure 7, two such plots, split by prominence condition, represent how listeners' preference changes over time and across the continuum, based on target word prominence. Note that color represents listeners' Elog-transformed preference for "ebb", over time. A value of zero (in the middle of the color scale) represents no preference, while a positive

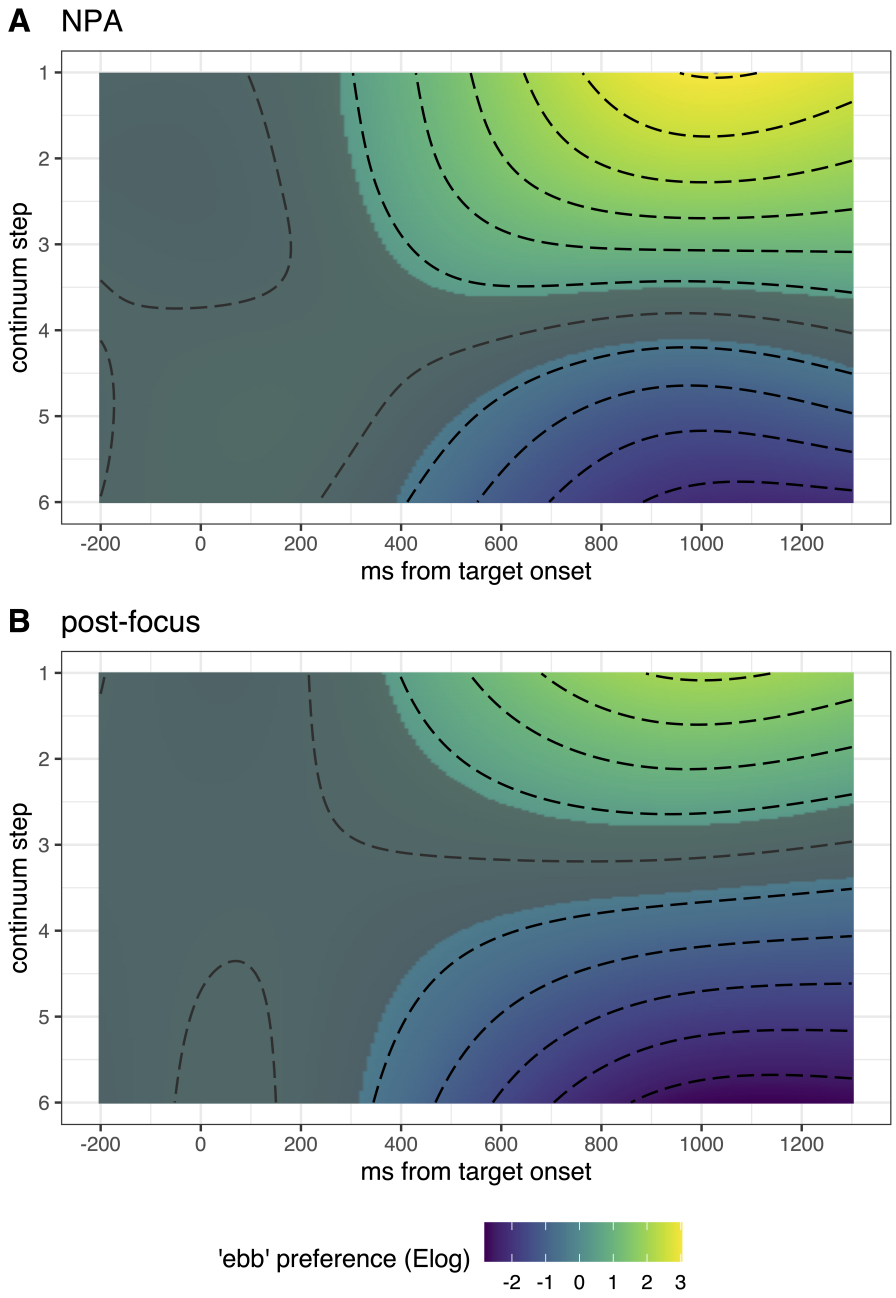

Figure 7. Topographic surface plots showing the effect of continuum step (y axis) over time (x axis), split by prominence condition. The color scale represents the degree of listeners' "ebb" preference. Shading on the surface (the darker color that covers the leftmost portion of the surface entirely) 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.

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". The shading represents a location on the surface 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 300ms). As time progresses, listeners develop graded preferences based on continuum step (as in Figure 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 ambiguous steps remain ambiguous even at the end of the analysis window, shown by the shaded area persisting until the end.

With this in mind, we now can assess the impact of prominence condition on listeners' use of the continuum over time (i.e. the non-linear interaction between (scaled) 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 shown 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 not surprising, given that we see divergence in looks based on prominence. 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 the shaded areas (where listeners do not have a significant preference for either target) differ in how they occupy space in the surface, and crucially, within a panel, which steps on the continuum show a preference first. 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. Notably, if prominence were *only* a later stage influence, we should expect the shape of the surfaces to be the same earlier in time. 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). That is, listeners looked to a target more quickly in the NPA condition, and ambiguity persists 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). [13]

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 piece of evidence that the prominence manipulation is shaping listeners' perception of formant cues directly.

As a way of synthesizing the two findings obtained from the divergence measure (Figure 6) and the surface plots (Figure 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

[13]Notably this asymmetry 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.

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 normalized effect estimates are shown in Figure 8, corresponding to the smooth differences shown in Figure 6.

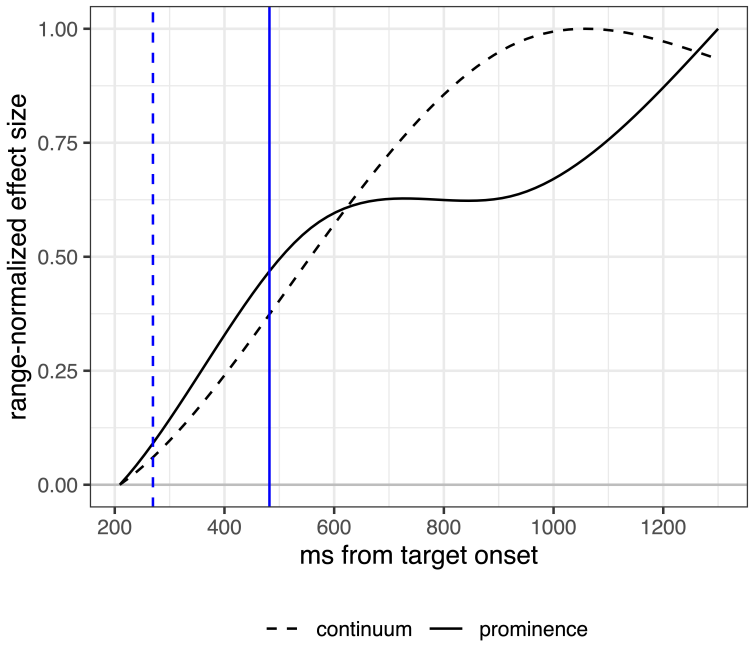

Figure 8. Range-normalized smooth differences, as a proportion of maximum difference, from 200 ms onward following the target onset, corresponding to the smooths in Figure 6. The solid line represents the effect of phrasal prominence, the dotted represents the effect of the continuum. Blue vertical lines indicate when an effect becomes significant as assessed by the smooth divergence analysis.

This normalized comparison allows us to inspect how each effect grows and changes over time, relative to its maximum and minimum (here 200 ms after target onset and onward, following Reinisch and Sjerps 2013). As can be seen in Figure 8, the effect of continuum grows steadily, peaking around 1000 ms following the target (cf. Reinisch and Sjerps 2013). The prominence effect is clearly different: it grows in tandem with the effect of the continuum, 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 contin-

uum effect (cf. Reinisch and 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.

This timecourse finding is similar to a pattern observed by Kim et al. (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). In showing the same general pattern as Kim et al. (2018), the present effect of prosodic prominence might be taken to reflect similar perceptual processing, discussed below.

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. We can therefore more generally say that the effect of prominence starts relatively early, while these early effects are subtler, and more robust prominence effects only occur later (as indexed by the divergence between smooths). 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 5: the differences between prominence conditions start early, and grow slowly over time to reach a relatively stable maximum, seemingly around 600 ms. As described above, Kim et al. (2018) observed subtle, non-significant effects of their AP-phrasing manipulation preceding the later stage (800 ms) point in time at which the effect became significant. Taken with the present results, this suggests that prosodic context effects in general might exert weak early influences in processing with stronger effects emerging later in processing.

5. Discussion

The two experiments presented here 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. This adjustment in categorization follows from sonority expansion (F1 raising/F2 lowering) in prominent vowels, showing that listeners account for how prominence modulates formant structure in vowel perception, for vowels like /ɛ/ and /æ/ which undergo sonority expansion.[14]

This finding is line with research that has tested how prosodic boundaries modulate perception of segmental contrasts (Katsuda & Steffman, 2020; Kim et al., 2018; Mitterer et al., 2019; Steffman, 2019b), and supports a model where listeners integrate prosodic context in perception. In testing phrasal prominence, Experiment 1 offered an extension of past studies, and predicts that further work looking at prominence marking 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 condition diverged from one another. This pattern, taken by itself, is wholly consistent with the prosodic analysis model proposed by Cho et al. (2007); Kim et al. (2018), wherein formant cues activate lexical hypotheses, and phrasal prosody is integrated later via lexical competition. However, an early influence of prominence (as shown in Figures 7 and 8), which shapes the earliest stages of formant use, suggests more nuance is needed.

[14] As noted in Section 2.1, not all vowels show clear sonority expansion effects in speech production (Cho, 2005; de Jong, 1995). For example prominent high vowels (e.g. /i/ in American English; Cho 2005) can show more extreme articulations under prominence (contra sonority expansion). Future work will accordingly benefit from testing how phrasal prominence impacts listeners' perception of other vowel contrasts, including those which do not undergo sonority expansion.

Recall the discussion in Section 1.2, which made a distinction between phonological prominence related to phrasal organization, and phonetic prominence, which might derive simply from relative acoustic salience in a given context. Though, importantly, 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 1). The opposite is true in the NPA condition. In this sense, listeners' perception of acoustic/phonetic prominence of the target likely varied across conditions. An immediate effect of prominence condition would accordingly reflect listeners' incorporation of acoustic 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., prosodic structural organization) but rather general 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. If we assume boundary processing is tightly tied into the overall prosodic organization of an utterance, following Cho et al. (2007) and Mitterer et al. (2019), we can see why it should necessitate prosodic analysis. Prominence perception, being multidimensional in nature (Baumann & Winter, 2018; Mo, 2011), is not strictly linked to the computation of phrasal prosody. [15] In this light we could take the effects seen in Experiment 2 as showing 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 which reinforces this effect, leading to more robust influences late in

[15] It is worth reiterating here however that Kim et al. (2018) found subtle, non-significant effects of AP phrasing on processing, which highlights the need to explore further how boundary processing might also exert early phonetic context effects.

processing.

Accordingly, a model of prominence and segmental processing must allow prominence driven re-tuning of perceived cue values to operate early in processing. The far reaching influence of phrasal prominence on segmental processing is analogous to that of prosodic structure on speech production according to some models (Keating & Shattuck-Hufnagel, 2002; Krivokapić, 2012, 2014), where prosodic structure plays a central role in organizing phonological and phonetic structure. If prosodic structure entails fine-grained modulations of acoustic features in speech (or, co-variance of acoustic properties in a systematic way), it is perhaps not surprising that prominence shapes listeners' perception of segmental cues prior to lexical access. It is proposed here that prominence effects should be particularly susceptible to these early perceptual influences, given the gradient and phonetically grounded nature of perceived prominence (Baumann & Cangemi, 2020). In this light, prominence may play a fundamentally different role in segmental processing as compared to prosodic boundaries. As discussed above, a prosodic boundary necessarily forms part of a hierarchically organized prosodic structure, and tend to be viewed as falling into fairly clean-cut categorical distinctions (Beckman and Pierrehumbert 1986; Carlson, Clifton, and Frazier 2001, though see Wagner and Crivellaro 2010). As such, prosodic boundary perception derives from a parsed prosodic structure, and therefore modulates segmental processing at a later stage, following Kim et al. (2018); Mitterer et al. (2019). Perceived prominence is not tethered to prosodic structure in the same way: it varies as a function of both phonological organization, phonetic cues, and various other pieces of information such as, for example, lexical frequency, part of speech and information structure (Baumann & Cangemi, 2020; Baumann & Winter, 2018; Bishop, 2012, 2017; Cole et al., 2019). The continuous and phonetically based nature of prominence perception is compatible with the view that it enters early into segmental processing, and as such, it is proposed here, differs in a key way from the use of prosodic boundary information.

The present findings have shown one novel way in which prosodic context influences spoken word recognition, shaping perception of formant cues even early in processing. Going forwards, future work will benefit from extending these results to explore how other prominence-lending contexts shape segmental processing. For exam-

ple, information structure influences prominence perception independent of acoustic information in the speech signal (Bishop, 2012). Seeing if semantic context which manipulates information structure leads to changes in segmental perception, and if so, what the timecourse of these effects is, would be informative. Based on the present results we can predict this sort of information structural effect would operate later in processing since it is not related to phonetic prominence in the speech signal. A further complementary test could focus on localized (or, segmental) prominence cues such as lengthened voice onset time in aspirated stops, nasal duration in nasals, or glottalization and laryngealized voice quality (e.g. Cho and Keating 2009; Cho, Kim, and Kim 2017; Dilley et al. 1996; Garellek 2014). These cues encode phrasal prominence in a systematic way (following e.g., Keating and Shattuck-Hufnagel 2002), and as such might be expected to signal prominence to listeners. Testing if localized prominence marking independently shapes vowel perception along these lines would help us understand exactly what acoustic features cue prominence in segmental processing. Together, these various lines of future work will give a better understanding of how prominence information more broadly construed is incorporated in processing, and how it shapes spoken language comprehension.

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.

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.

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., & Winter, B. (2018). What makes a word prominent? predicting untrained

german listeners' perceptual judgments. *Journal of Phonetics*, *70*, 20–38.

Beckman, M. E., & Ayers, G. (1997). Guidelines for ToBI labelling. *The OSU Research Foundation*, *3*, 30.

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

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 jason bishop. *Prosody and meaning*, *25*, 239.

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.

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., 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.

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

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.

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

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. Retrieved from `http://onlinelibrary.wiley.com/doi/10.1111/lnc3.12178/abstract`

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., 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. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0095447006000167`

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

Chong, J., & Garellek, M. (2018). Online perception of glottalized coda stops in american english. *Laboratory Phonology*.

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. Retrieved 2019-09-30, from `https://asa.scitation.org/doi/10.1121/1.2000774`

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., 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*, *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).

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.

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

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.

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.

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

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

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.

Fougeron, C., & Keating, P. (1997). Articulatory strengthening at edges of prosodic domains.

Journal of the Acoustical Society of America, *106*(6), 3728–3740.

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

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

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. Retrieved 2019-09-30, from `https://asa.scitation.org/doi/abs/10.1121/1.419712`

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.

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.

Katsuda, H., & Steffman, J. (2020). Intonational cues to prosodic boundary influence perception of contrastive vowel length in Tokyo Japanese. In *Proceedings of the 10th international conference on speech prosody, tokyo, japan* (pp. 56–60). Retrieved from `http://dx.doi.org/10.21437/SpeechProsody.2020-12`

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. *Speech production: Models, phonetic processes, and techniques*, 167–186.

Keating, P., Cho, T., Fougeron, C., & Hsu, C.-S. (2004). Domain-initial articulatory strengthening in four languages. *Phonetic interpretation: Papers in laboratory phonology VI*, 143–161.

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

Kim, S., & Cho, T. (2013). Prosodic boundary information modulates phonetic categorization.

The Journal of the Acoustical Society of America, *134*(1), EL19–EL25. Retrieved from http://scitation.aip.org/content/asa/journal/jasa/134/1/10.1121/1.4807431

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.

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

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.

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.

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

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

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

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.

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

McMurray, B., & Jongman, A. (2011). What information is necessary for speech categoriza-

tion? harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological review*, *118*(2), 219.

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

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, campinas, brazil* (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.

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* (Vol. 28). Walter de Gruyter.

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.

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. Retrieved 2019-09-30, from https://asa.scitation.org/doi/abs/10.1121/1.1906875

Pierrehumbert, J. B. (1980). *The phonology and phonetics of english intonation* (Unpublished doctoral dissertation). Massachusetts Institute of Technology.

Pierrehumbert, J. B., & Talkin, D. (1992). Lenition of/h/and glottal stop. *Papers in laboratory phonology II: Gesture, segment, prosody*, *90*, 117.

Plantinga, J., & Trainor, L. J. (2005). Memory for melody: Infants use a relative pitch code. *Cognition*, *98*(1), 1–11.

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.

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. (1997). Spectral envelope and context effects in the tritone paradox. *Perception*, *26*(5), 645–665.

Roessig, S., Mücke, D., & Pagel, L. (2019). Dimensions of prosodic prominence in an attractor model. *Proc. Interspeech 2019*, 2533–2537.

Schellenberg, E. G., & Trehub, S. E. (2003). Good pitch memory is widespread. *Psychological Science*, *14*(3), 262–266.

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.

Sóskuthy, M. (2017). Generalised additive mixed models for dynamic analysis in linguistics: a practical introduction. *arXiv preprint arXiv:1703.05339*.

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.

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.

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.

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.

Umeda, N. (1975). Vowel duration in american english. *The Journal of the Acoustical Society of America*, *58*(2), 434–445.

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. Retrieved 2019-09-29, from `https://asa.scitation.org/doi/abs/10.1121/1.395284`

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, chicago, il.*

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. glasgow, scotland.*

Winn, M. (2016). Vowel formant continua from modified natural speech (Praat script). `http://www.mattwinn.com/praat/Make_Formant_Continuum_v38.txt` [Computer software manual]. 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.

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). Retrieved 2019-09-30, from `https://escholarship.org/uc/item/2pm9c9sq`

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.