Vowel-initial glottalization as a prominence cue in speech perception and online processing

Jeremy Steffman
Northwestern University
jeremy.steffman@northwestern.edu

word count: 10,821

Abstract

Three experiments examined the relevance of vowel-initial glottalization in the perception of vowel contrasts in American English, in light of the claimed prominence-marking function of glottalization in word-initial vowels. Experiment 1 showed that the presence of a preceding glottal stop leads listeners to re-calibrate their perception of a vowel contrast in line with the prominence-driven modulation of vowel formants. Experiment 2 manipulated cues to glottalization along a continuum and found that subtler cues generate the same effect, with bigger perceptual shifts as glottalization cues increase in strength. Experiment 3 examined the time-course of this effect in a visual world eyetracking task, finding a rapid influence of glottalization which is simultaneous with the influence of formant cues in online processing. Results are discussed in terms of the importance of phonetically detailed prominence marking in speech perception, and implications for models of processing which consider segmental and prosodic information jointly.

keywords: speech perception, prominence, glottalization, vowels, eyetracking

1 Background

One important question in prosody research is the following: How do speakers make syllables and words prominent in speech, and how do listeners make use of this information? The answer to this question is complex, entailing a consideration of a language's various cues to prominence, and the listener's incorporation of prominence information in different domains of perception and processing.

In speech production, the literature has documented various ways in which speech articulations and acoustics are modulated by prosodic prominence, referred to here under the umbrella term of "prominence strengthening". These effects generally help enhance a given segment's perceptual salience, and/or enhance acoustic (or featural) properties relevant for the contrast system of a given language (e.g., Cho, 2005; Garellek, 2014; Cole et al., 2007; de Jong, 1995; Beckman et al., 1992; Kim et al., 2018a).

In comparison, relatively little work has been carried out examining the perceptual component of the above question. The present study thus addresses one part of this line of inquiry from the perspective of the listener. In three experiments, this study tests how glottalized voice quality and production of a glottal stop impact the perception of vowels in American English, in line with the hypothesized function of glottalization as prominence marking. The perception of /ɛ/ versus /æ/ is adopted as a test case. A visual-world eyetracking experiment further tests how the influence of glottalization plays out in online speech processing, informing our understanding of how prominence cues are integrated as speech unfolds.

1.1 Vowel-initial glottalization in speech production

Glottalization is used here as a cover term to refer to the production of a sustained closure of the vocal folds, i.e. a glottal stop [ʔ], and localized voice quality changes that are associated with constriction of the vocal folds during voicing (Garellek, 2013; Huffman, 2005). The cover term is useful if we consider the latter of these to be an "incomplete" or lenited glottal stop realization, as is common in the literature (Pierrehumbert and Talkin, 1992; Dilley et al., 1996).

Of the languages described in the UPSID database (Maddieson and Precoda, 1989), about half use glottalization contrastively, and in these languages /ʔ/ can be considered a phoneme. However, in many languages that do not use glottalization contrastively, it is well documented that glottalization is nevertheless pervasive in speech, for example in English, Dutch, and Spanish (Dilley et al., 1996; Jongenburger and van Heuven, 1991; Garellek, 2014). An important task for speech research is thus accounting for the prevalence and distribution of glottalization in spoken language.

One clear predictor of glottalization in American English (among other languages) is prosodic organization. Glottal stops are described as being “inserted” at the beginning of vowel-initial words in prosodically strong positions. Prosodically strong positions include the beginning of a prosodic phrase (Pierrehumbert and Talkin, 1992; Dilley et al., 1996), and in words which bear phrasal prominence (Dilley et al., 1996; Garellek, 2013). Dilley et al. (1996) in particular show that phrase-medial, word-initial vowels in pitch-accented (phrasally-prominent) syllables are glottalized at higher rates as compared to non-prominent equivalents. Notably however, not all pitch accented word-initial vowels are glottalized, and vowels in words which lack pitch-accent but do not have a reduced vowel are more likely to be glottalized than reduced vowels. Speakers also vary widely in their overall rate of glottalization and the extent to which prominence impacts their rate of glottalization. In this sense, glottalization in word-initial vowels is only probabilistically related to prominence marking, though with a robust tendency to co-occur with phrasal prominence.

Garellek (2013, 2014) further suggests a tight link between prominence marking and glottalization in American English, using electroglottography (EGG) to examine voicing in vowel-initial words. Garellek (2014) found that phrase-initial vowels, especially non-prominent vowels, were generally produced with less vocal fold contact during voicing, corresponding to breathy voicing. This effect also became larger at higher-level phrasal domains. Breathier phrase-initial voicing was attributed to phrase-initial pitch reset, where falling pitch (immediately after reset) results in relaxation of the cricothyroid and thyroarytenoid muscles, and vocal fold abduction (Mendelsohn and Zhang, 2011; Zhang, 2011). Breathier voicing generally leads to decreased intensity and weaker formant energy (Garellek and Keating, 2011; Gordon and Ladefoged, 2001), and Garellek (2014) accordingly proposes that phrase-initial glottalization occurs as a countervailing influence which mitigates the effects of pitch-reset-induced breathiness on voice quality, explaining its prevalence phrase-initially (Pierrehumbert and Talkin, 1992; Dilley et al., 1996). Glottalization in prominent phrase-initial vowels “strengthens” voice quality, as described by Garellek, in the sense that it engenders more high frequency energy and overall intensity, and boosts frequency information that will be useful in vowel perception (Kreiman and Sidtis, 2011; cf. Garellek, 2013 who found a boost of harmonic energy between 1500 - 2500 Hz). Glottalization may also be functionally useful in prominence-marking in separating prominent vowel-initial words from surrounding material, and modulating the amplitude envelope in the vicinity of prominent vowels to make them stand out. Preceding silence from a glottal stop will likewise give a boost to listeners' auditory system at the onset of the vowel (Delgutte, 1980; Delgutte and Kiang, 1984). This view of phrase-initial (and phrase-medial) glottalization implicates prominence as the driving force behind it. In this sense glottalization in word initial-vowels in American English is an example of phonetic prominence strengthening, though perceptual evidence along these lines is currently lacking.

1.2 Prosody and prominence in segmental perception

The role of prosodic features such as prominence in segmental perception and lexical processing has been a recent topic of interest in the literature (Mitterer et al., 2016; Kim et al., 2018b; Mitterer et al., 2019; McQueen and Dilley, 2020). As such, in addition to contributing to our understanding of glottalization as a perceptual prominence cue, the present study will build on our understanding of the role of prominence (and prosody more generally) in segmental perception and spoken word recognition.

As alluded to above, it is well documented in the speech production literature that prosodic organization modulates cues that are relevant in the perception of segmental contrasts (see e.g., Keating, 2006 for an overview). For example, voice onset time (VOT) in aspirated stops, an important cue for voicing contrasts, varies systematically as a function of prosodic factors. VOT is longer at the beginning of prosodic domains and in phrasally prominent positions (Cole et al., 2007; Keating et al., 2004; Kim et al., 2018b). Another example of prosodically modulated cues to segmental contrasts, described in more detail in Section 1.3, is that of vowel formants.

To the extent that phrasal prosody impacts segmental realization along these lines, the listener is hypothesized to benefit from integrating prosodic information with their perception of segmental and lexical material (Kim and Cho, 2013; Mitterer et al., 2016). A model which has framed this line of inquiry and received clear empirical support is that of *prosodic analysis* (Cho et al., 2007; McQueen and Dilley, 2020). The model's architecture stipulates simultaneous parses of segmental information and prosodic information from the speech signal, though the role of each of these in processing are different. Adopting an activation-competition view of word recognition, the model postulates that segmental information activates entries in the lexicon, while phrasal prosodic information is used to select among possible candidates. In the original formulation of the model this entails the reconciliation of prosodic boundaries and word boundaries to determine lexical selection (cf. Christophe et al., 2004). Empirical support for the model comes from studies showing a delayed influence of prosodic boundary information in processing (Kim et al., 2018b; Mitterer et al., 2019), consistent with a post-lexical influence in word recognition. What the model and current data show more generally is the importance of considering both prosodic and segmental factors as being processed in parallel in speech recognition (see McQueen and Dilley, 2020 for a recent overview).

With respect to glottalization specifically, recent perception and processing studies in Maltese, a language in which /ʔ/ is contrastive, suggest that listeners are highly sensitive to its prosodic patterning in the language (Mitterer et al., 2021a, 2019, 2021b). In addition to marking a phonemic contrast in Maltese, vowel-initial words can be glottalized when they are at the beginning of a prosodic phrase as a form of phrase-initial strengthening. Glottalization thus serves a sort of dual function, it is phonemic and conveys contrast, and also patterns based on prosodic organization. Mitterer et al. (2019) show that listeners are aware of this dual patterning: when a word is phrase-initial, the listener is more likely to attribute the presence of glottalization as being driven by prosody, thus inferring a phonemically vowel-initial word. In contrast, when glottalization precedes a vowel phrase-medially, the listener is more likely to infer that the word is phonemically/contrastively glottalized. Consistent with the prosodic analysis model, these effects were seen to be delayed in time, as assessed in a visual world eye-tracking study. Mitterer et al. (2021a) show that glottal stops differ from other stops (e.g., /t/) in that they do not strongly constrain lexical access, suggesting that listeners' interpretation of glottalization is intimately linked to prosodic features in a way that differs from other stops. Mitterer et al. (2021b) further show that glottalization is clearly interpreted as a prosodic feature in that it impacts syntactic parsing decisions in the resolution of attachment ambiguity: the presence of word-initial glottalization leads listeners to posit a preceding prosodic boundary, and thus the presence of a syntactic boundary. These results together thus suggest that vowel-initial glottalization can be treated as prosodic cue in perception by listeners, even when glottalization is contrastive.

Steffman (2021a) offers another relevant comparison for the present study. Steffman examined the influence of prosodic prominence, as cued by the intonational tune and durational patterns of a phrase, on listeners' perception of vowel contrasts. Vowels are strengthened phonetically by formant modulations described in Section 1.3 below. Steffman thus tested how phrase-level prominence impacted the perception of vowel formants, and further examined the timecourse of its influence. In contrast to the strictly delayed influence of prosodic boundaries documented in previous studies (Kim et al., 2018a; Mitterer et al., 2019), Steffman found that phrasal prominence showed subtle earlier influences in vowel perception, though these effects were quite small, and strengthened over time to be most robust later in processing. This suggests prominence may modulate earlier sublexical processing (in shaping the mapping of vowel acoustics to segmental representations), as the listener determines if an individual segment has undergone prominence strengthening. Prominence is however strongest in its influence at a later stage, consistent with a parse of prominence in a more global prosodic structure (e.g., the presence or absence of pitch-accentuation) being reconciled with lexical candidates, under the hypothesis that the lexicon contains information about prosodically conditioned pronunciation variants along the lines of Brand and Ernestus (2018); Pitt (2009); Mitterer et al. (2021a). Notably, Steffman (2021a) only considers prominence as a more global and phrasal context.

1.3 The present study

Given these recent studies on the role of prominence in vowel perception and the processing of vowel-initial glottalization, the present experiments will inform if prominence cued by glottalization should be considered as a mediating factor in segmental perception in American English, a language where glottalization is not contrastive. To the extent that vowel-initial glottalization is a relevant prominence cue, we can examine the timecourse of its influence in relation to the general prediction from the prosodic analysis model that prosody shows a delayed influence in processing. We can further consider the present results in relation to the data from Steffman (2021a), as a comparison of more global (phrasal) prominence and a highly local prominence cue (glottalization).

Relevant to the present study, the literature documents a variety of ways in which vowel articulations may be modulated under prominence. Typically, prosodic prominence is here considered in terms of phrase-level prominence marking: the presence/absence of a pitch accent on a syllable. A well-documented pattern of prominence strengthening in vowels has been termed *sonority expansion*, where sonority is defined as "the overall openness of the vocal tract or the impedance looking forward from the glottis" (Silverman and Pierrehumbert, 1990, p 75). In this sense, a more sonorous vowel articulation is one which is produced with increased amplitude of jaw movement and other articulatory adjustments that allow more energy to radiate from the mouth. Sonority-expanding gestures make a vowel articulation more acoustically prominent (louder, longer etc.), and have been described as enhancing its "sonority features" (de Jong, 1995). Other effects, not consistent with sonority expansion, have also been documented in the literature, for example, the production of more extreme high vowel articulations (as with /i/), which are not more open but instead reflect hyperarticulation of the vowel target under prominence (Cho, 2005; Erickson, 2002; de Jong, 1995). In this sense, patterns of prominence strengthening are dependent on the vowels under consideration, and the system of contrasts in the language (e.g., Cho, 2005; Garellek and White, 2015), and so is the listener's perception of vowels a function of prominence (Steffman, 2020).

Vowels which *do* undergo sonority expansion are realized as acoustically lower and backer in the vowel space, with higher F1 and lower F2 (Cho, 2005), and listeners' perception of prominence in a prominence rating task reflects this formant variation as well (Mo et al., 2009). This pattern will form the basis of the test case adopted in the present study as we ask if listeners expect a more prominent variant of a vowel (specifically with higher F1 and lower F2) to be realized in a prominent context.

These questions raised in Section 1 are addressed in testing if a glottal stop modulates vowel perception in line with sonority expansion effects on vowel formants (Experiment 1), using the contrast between /ɛ/ and /æ/ as a test case (vowels which undergo sonority expansion). This study further tests if fine-grained glottalization cues that do not entail a sustained stop generate the same effect (Experiment 2), and if glottalization mediates online processing of vowel information in the ways predicted by the current model of prosodic analysis (Experiment 3). The experiments consist of an offline two-alternative forced choice task, and a visual world eyetracking task, in which listeners categorized stimuli on an /ɛ/-/æ/ continuum with various contextual manipulations of glottalization. All of the stimuli used in the present experiments, the data for each experiment, and the scripts used the analyze the data are included in full in the open-access repository for the paper hosted on the OSF at https://osf.io/v4cdz/.

2 Materials

The materials used in all experiments reported here were created by re-synthesizing the speech of a male American English speaker. The speech material was recorded in a sound-attenuated booth in the UCLA Phonetics Lab, using an SM10A Shure™ microphone and headset. Recordings were digitized at 32 bit with a 44.1 kHz sampling rate.

2.1 A full glottal stop: Experiments 1 and 3

The goal in creating the stimuli was to design a continuum that varied in F1 and F2, ranging between two vowels, and manipulate the presence or absence of preceding glottalization. The two words used as endpoints of the continuum were "ebb" /ɛ/, and "ab" /æ/. F1 and F2 were manipulated by LPC decomposition and resynthesis using the Burg method (Winn, 2016) in

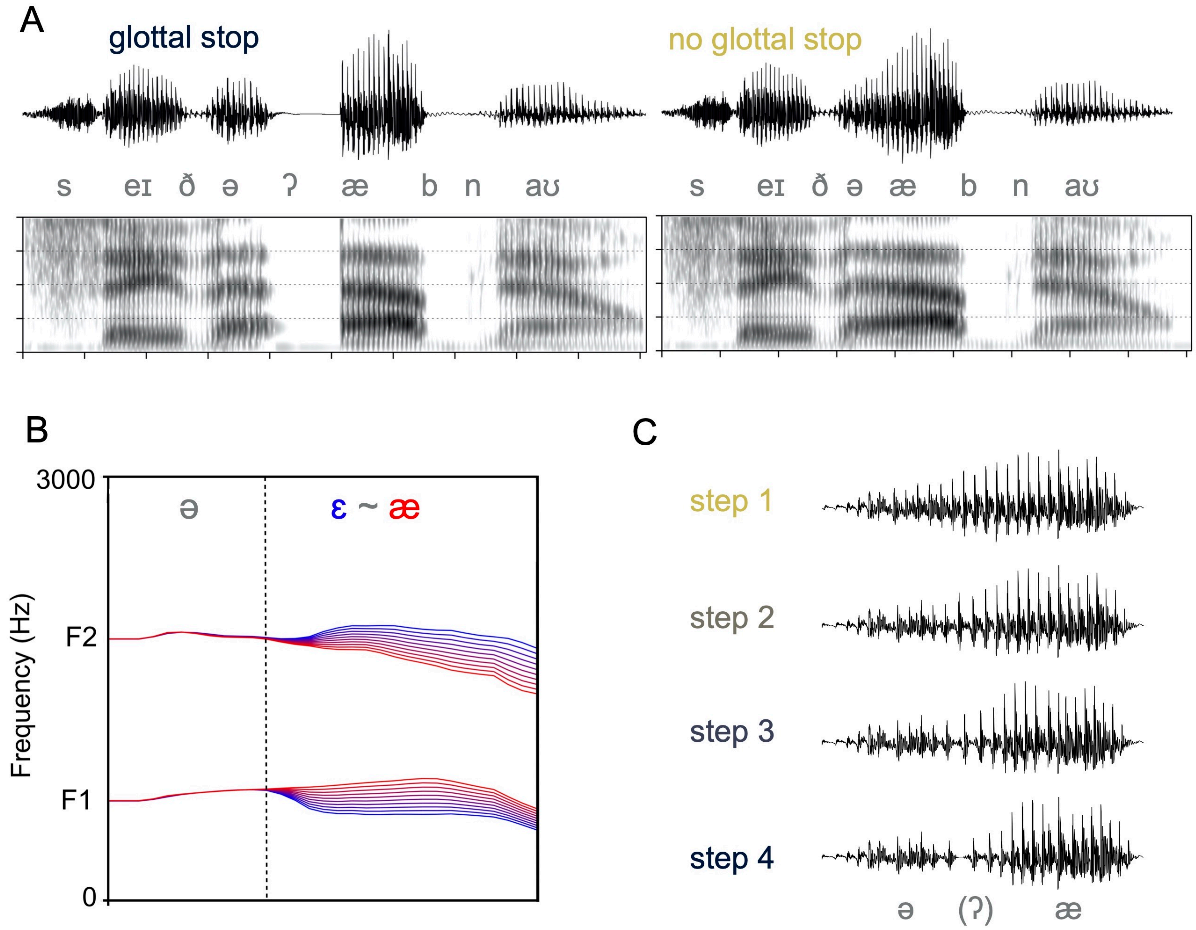

Figure 1: Visualizations of the stimuli used in all Experiments. Panel A: waveforms and spectrograms showing the glottal stop manipulation (y axis 0-4000 Hz, ticks on axis at 100 ms intervals; in this example the target vowel is at step 10, the most /æ/-like). Panel B: formant tracks showing the 10-step continuum created from the VV sequence (the target and the preceding vowel). Panel C: Waveforms showing the four steps of the glottalization continuum from Experiment 2, with just the target vowel and preceding vowel shown.

Praat (Boersma and Weenink, 2020). The formant values for each endpoint were based on model sound productions of “ebb” and “ab”. The resynthesis process estimated the 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 space (Traunmüller, 1990). Phase-locked higher frequencies from the starting base /ɛ/ model were restored to all continuum steps, improving the naturalness of the continuum. The result was a 10 step continuum ranging from /ɛ/ to /æ/ values in F1 and F2. Intensity and pitch were invariant across the continuum.

The starting point for stimulus creation was a production of the sentence “say the ebb now”, with “the” produced as [ðə]. The sentence was produced with pitch accents on the word “say” and “ebb”, such that the word with the target vowel bore the final (nuclear) pitch accent in the phrase. The creation of the continuum only altered F1 and F2 in the target word as described above, creating a [ðəɛb] to [ðəæb] continuum, with continuous formant transitions from the precursor vowel to the target. Formant tracks for the 10-step continuum, and preceding vowel are shown in Figure 1 panel B. This constitutes what will be referred to as the “no glottal stop condition”, where no glottal stop preceded the target sound in the vowel hiatus environment. The formants in the precursor vowel [ə] were also slightly centralized (F1 raised, F2 lowered) so that these manipulations were not confounded with spectral contrast effects.[1] This manipulation made the precursor vowel sound slightly lower than a canonical [ə], though it was clearly intelligible and judged to sound natural.

The goal in creating the “glottal stop condition” was to cross-splice [ʔ] from a different production of the carrier phrase in which it preceded the target. The portion of the glottal stop that was inserted was the silent closure (approximately 100 ms in duration), and the short aperiodic burst that accompanied the release of the stop (approximately 15 ms). The production from which [ʔ] was cross-spliced was [ðəʔæb]. In the case that any information about the following vowel is contained in the release of the stop (though none was perceived), it would bias listeners towards /æ/ when a glottal stop precedes the target, which is the opposite of the predicted prominence effect, described in Section 3. The point at which the glottal stop was inserted was where formant trajectories began to shift to the target vowel, indicated by the dashed vertical line in Figure 1, panel B. The insertion of [ʔ] resulted in a sudden end to the vowel in the precursor. To render the precursor more natural, several periods from [ə] in the production of [ðəʔæb] were cross-spliced and appended to the end of the precursor vowel at a zero crossing in the waveform. These added periods did not replace any portion of the precursor vowel but instead introduced a dip in amplitude and irregular voicing going into the glottal stop, which was judged to improved the naturalness of the stimuli. This modified precursor vowel and following [ʔ] were cross-spliced to precede all steps on the continuum, resulting in a [ðəʔɛb] to [ðəʔæb] continuum, as shown in Figure 1 panel A.

2.2 A glottalization continuum: Experiment 2

As is well documented in the speech production literature, and noted above, the way in which glottalization is realized phonetically is notoriously variable, and needn't entail the production of a sustained stop at the glottis (Garellek, 2013; Dilley et al., 1996). As such, an important question is if different realizations of a glottal stop produce similar perceptual effects. Various studies have shown that glottalization may be cued perceptually by changes in pitch and intensity (Gerfen and Baker, 2005; Pierrehumbert and Frisch, 1997). Accordingly, Experiment 2 was designed to create a continuum that varied in glottalization strength. Step 1 in the glottalization continuum in Experiment 2 was the same as the "no glottal stop condition" in Experiment 1. Three additional glottalization conditions were created (labeled step 2-4 in Figure 1C). In each, pitch and intensity cues were varied to signal an increase in the strength of glottalization between the pre-target and target vowels.

This was accomplished by decreasing the f0 and intensity at the juncture of the two vowels, indicated by the dashed vertical line in Figure 1 panel B. The seven f0 periods including and surrounding this point were manipulated. Intensity was manipulated as a 2 dB decrease in intensity per glottalization continuum step for these seven periods, which were then cross-spliced into the original unmodified production at zero crossings in the waveform. The pitch manipulation, which was implemented with the PSOLA method in Praat (Moulines and Charpentier, 1990) took the f0 period at the juncture and decreased it linearly by 25 Hz at each step. An original f0 of approximately 115 Hz at Step 1 thus became 90, 65, and 40 Hz at Steps 2,3 and 4 respectively. f0 was interpolated linearly from this low point across the surrounding three periods on either side to the f0 values surrounding them. The result was a continuum in strength of glottalization, shown in Figure 1 panel C.

Experiment 2 used a subset of the formant continuum steps from Experiment 1, as it was observed that listeners in Experiment 1 were essentially at ceiling in their categorization responses for steps 1-3. For this reason only steps 3-10 from Experiment 1 were used, though they are renumbered as steps 1-8 in presenting the results (step 1 in Experiment 2 corresponding to step 3 in Experiment 1, and so on).

3 Experiment 1 and 2

Experiment 1 and 2 are presented together here, given their similarity. Let us first consider the empirical predictions that we would expect if glottalization cues prominence to listeners and exerts an influence on their perception of the vowel contrast. If a vowel preceded by glottalization is perceived as prominent, a more prominent acoustic realization of that vowel may be expected by listeners. In this case, it would mean a lower and backer realization of the vowel (with higher F1 and lower F2), with a prominent /ɛ/ essentially becoming acoustically more like /æ/. The corresponding perceptual response would thus be a shift in categorization of the F1/F2 continuum, with more sonorant (lower, backer) F1/F2 values categorized as /ɛ/ in a prominent context (when preceded by glottalization), as compared to a non-prominent one. Empirically, this predicts increased /ɛ/ responses under prominence. Such an effect would constitute perceptual re-calibration for a prominent vowel realization. It is worth noting here that Steffman (2021a) found this effect with the same contrast, when prominence was cued by global/phrasal context.

In Experiment 2, we can further predict that increasing strength of glottalization should entail increasing strength of this effect, where we see additive shifts in categorization from Steps 1 to 4 in the glottalization continuum shown in Figure 2C.

3.1 Participants and procedure

3.1.1 Experiment 1

30 participants were recruited for Experiment 1. All participants were self-reported native American English speakers with normal hearing, and were recruited from the student population at the University of California, Los Angeles. Each participant completed a language background questionnaire and provided informed consent to participate. Participants received course credit for their participation. The online platform that was used to control stimulus presentation was Appsobabble (Tehrani, 2020).

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

Participants were instructed that their task was to identify which word they heard by key press, where a "j" key press indicated the word on the right side 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 × 10 repetitions). Halfway through the test trials, participants were prompted to take a short self-paced break. The experiment took approximately 15-20 minutes to complete in total.

3.1.2 Experiment 2

34 participants were recruited from the same population for Experiment 2. Data collection and recruitment took place remotely due to COVID 19. Participants were asked to complete the experiment in a quiet location while using headphones. There were a total of 32 unique stimuli used in the experiment (8 formant continuum steps × 4 glottalization continuum steps) each of which was repeated a total of 7 times for a total of 224 trials in the experiment. The experimental procedure was otherwise the same as in Experiment 1.

3.2 Analysis

The analysis of categorization data in all experiments reported here was carried out with a logistic mixed-effects regression model, using a Bayesian implementation with the R package *brms* (Bürkner, 2017). The models were run using R version 4.1.2 (R Core Team, 2021) in the RStudio environment (RStudio Team, 2021). Weakly informative normally distributed priors were employed for both the intercept and fixed effects, as Normal(mean = 0, standard deviation = 1.5) in log-odds space.[2] The model was fit to draw 4,000 samples from the posterior in each of four Markov chains, with a burn-in period of 1,000 iterations in each chain. $\hat{R}$ and Bulk and Tail ESS were inspected to confirm convergence and adequate sampling.

In reporting effects on categorization two measures are given, both characterizing the estimated posterior distribution for a given fixed effect. First we report the estimate and 95% credible intervals (CrI) for an estimate. This gives the effect size (in log-odds), and characterizes the distribution/certainly around the estimate. When 95% credible intervals exclude 0, this suggests a consistently estimated directionality, and accordingly a robust influence. In comparison, 95% credible intervals which *include* 0 would indicate substantial variability in the estimated direction of an effect, and therefore a non-reliable impact on categorization. An additional metric is reported: the "probability of direction", (henceforth pd), computed with *bayestestR* package (Makowski et al., 2019). This metric is useful in that it corresponds more intuitively to a frequentist model's p-value. pd indexes the percentage of a posterior distribution which shows a given sign, and ranges between 50 and 100. A posterior centered precisely on zero (i.e, no effect), will have a pd of 50, while a posterior with a strongly skewed negative or positive distribution will have pd that approaches 100. Convincing evidence for an effect would come from pd values that are greater than 95. Tables showing all fixed effects estimates for each model are included in the appendix.

The model was coded to predict categorization responses, with an /ɛ/ response mapped to 1, and an /æ/ response mapped to 0. The formant continuum was coded as a continuous variable, and scaled and centered. In Experiment 1, glottalization was contrast coded with the presence of a glottal stop mapped to 0.5, and the absence of a glottal stop mapped to -0.5. Categorization responses were predicted as a function of continuum step, glottalization, and the interaction of these two fixed effects. In Experiment 2, the glottalization continuum was treated as a continuous variable, and was scaled and centered. Categorization responses were predicted as a function of glottalization continuum, formant continuum, and their interaction. Random effects in the each model included random intercepts for participant and random slopes for all fixed effects and interactions.

3.3 Results and discussion

The results of Experiments 1 and 2 are shown together in Figure 2. In both Experiment 1 ($\beta = -2.95$, 95%CrI $= [-3.27, -2.64]$; pd $=$ 100) and Experiment 2 ($\beta = -2.61$, 95%CrI $= [-2.99, -2.26]$; pd $=$ 100) changing formant values along the continuum shifted categorization in the expected way; increasing (scaled) step values along the continuum decreased the log-odds of an /ɛ/ response.

In Experiment 1, the glottal stop condition showed a credible effect in shifting categorization ($\beta = 1.69$, 95%CrI $= [1.26, 2.12]$; pd $=$ 100). As shown in Figure 1A, a preceding glottal stop increased /ɛ/ responses. This result lines up with the predictions outlined in Section 3, suggesting that listeners do indeed adjust their perception of the contrast in line with sonority expansion: a vowel preceded by a glottal stop is expected to be realized as a more prominent variant, i.e. lower and backer in the vowel space.

In Experiment 2, the glottalization continuum additionally showed a credible effect in shifting categorization responses ($\beta = 1.69$, 95%CrI $= [1.26, 2.12]$; pd $=$ 100). This is evident in Figure 2B as increasing rightward shifts along the glottalization continuum, with the strongest glottalization cues (step 4), showing the largest difference from step 1 (no glottalization). The

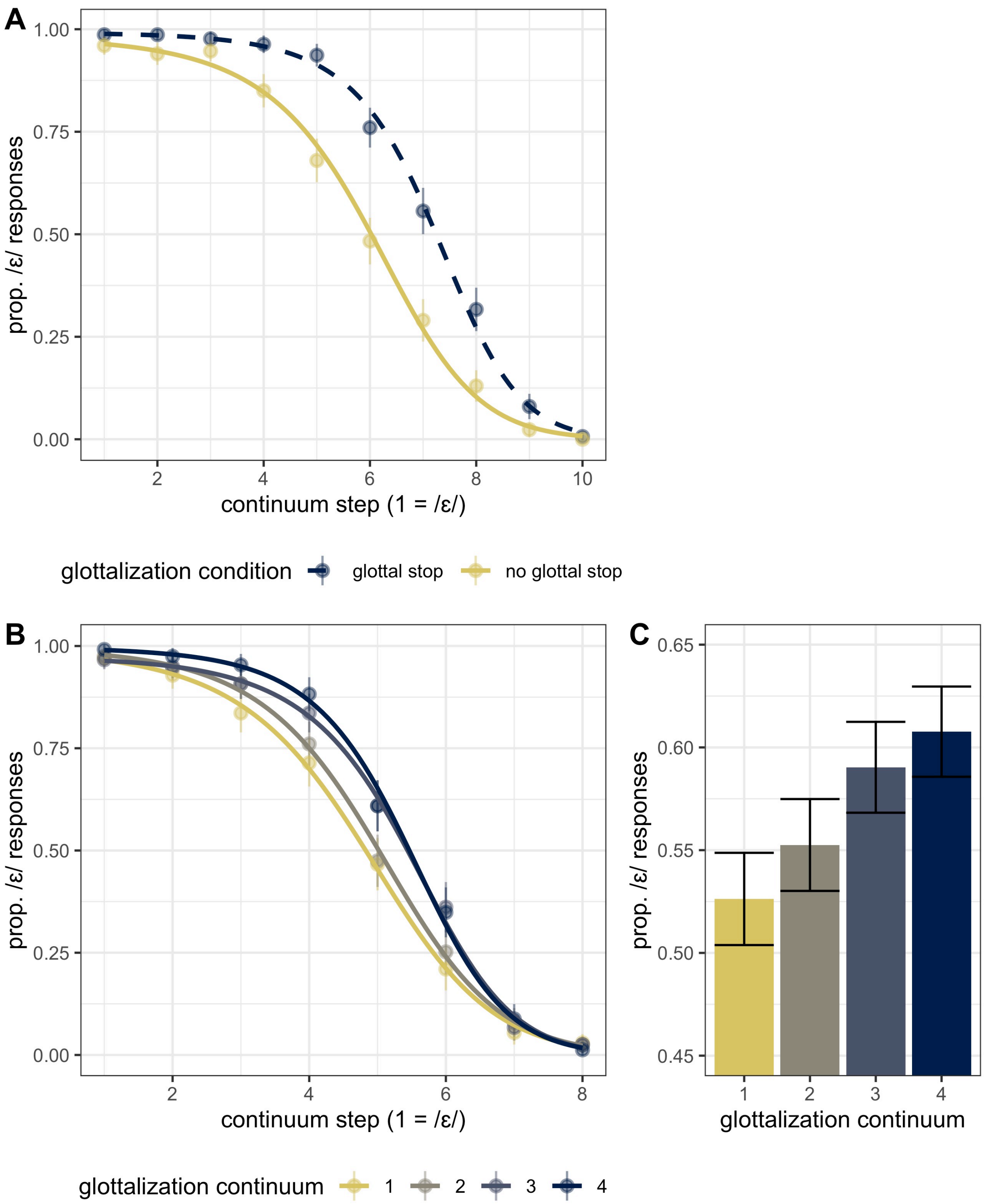

Figure 2: Categorization results in Experiment 1 (panel A) and 2 (panel B and C). In panels A and B, the x axis shows the formant continuum and the y axis shows listeners' proportion of /ɛ/, responses at each step, split by glottalization condition. Lines in panel A and B show a logistic fit to the data with points showing empirical means. Error bars show one SE from the data (not model estimates). Panel C shows the effect of the glottalization continuum on the x axis, pooled across formant continuum steps.

results are further shown in Figure 3B, which collapses across all steps of the formant continuum, showing a graded increase in /ɛ/ responses as glottalization cues increase in strength. In an alternative parameterization of the Experiment 2 model (included in the open-access repository for the paper but not reported here), the glottal stop continuum was treated as a categorical variable with four levels. In that model, pairwise comparisons between all levels, compared with *emmeans* (Lenth et al., 2018) were reliably different (all having pd $>$ 98).

The data from Experiments 1 and 2 thus supports the prediction that vowel-initial glottalization serves a prominence-marking function for listeners. Importantly, we also see that various different realizations of glottalization engender similar perceptual effects. Moreover, we see a detailed perceptual relationship between strength of glottalization and perceptual adjustments in the perception of formant cues. These effects also mirror the influence of phrasal prominence documented in Steffman (2021a), in other words, a phrasally prominent vowel is subject to the same perceptual shifts in categorization as a vowel preceded by glottalization.

4 Experiment 3

Given the clear effect of glottalization on categorization evident in both Experiment 1 and 2, Experiment 3 examined the timecourse of its influence in online processing in a visual world eyetracking task.

4.1 Materials

Experiment 3 made use of the same materials as Experiment 1, though it used a subset of the 10 step continuum. The method by which the Experiment 3 stimuli were selected was the same as that used in Mitterer and Reinisch (2013). The overall interpolated categorization function for Experiment 1 was inspected. The point at which the interpolated function crossed 50% (i.e. the most ambiguous region in the continuum) was identified. The three steps on each side of this crossover point were used in Experiment 3. This led to the selection of steps 4-9 from Experiment 1. Note that these steps are re-numbered as steps 1-6 in what follows, where step 1 in Experiment 3 refers to step 4 in Experiment 1, and so on. There were accordingly 12 unique stimuli used (6 continuum steps $\times$ 2 prominence conditions).

4.2 Participants and procedure

40 participants were recruited from the same population as previous experiments to participate in Experiment 3. Testing was carried out in a sound-attenuated room in the UCLA Phonetics Lab.

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

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

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

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

4.3 Eyetracking analyses

Two complementary analyses of the eyetracking data are presented here. The dependent measure in each analysis was a "preference measure", which offers a normalized measure of listeners' propensity to fixate on a target (cf. Reinisch and Sjerps, 2013). This measure is computed as log-transformed looks to "ebb" minus log-transformed looks to "ab", using the empirical logit (Elog) transformation[3] given in Barr (2008). In both analyses the analysis window of 0-1200 ms from the onset of the target vowel is used.

The first analysis was a traditional moving window analysis, which assesses how vowel-internal formant cues influence eye movements in relation to the glottal stop manipulation. Time bins of 100 ms were used with the preference measure computed at 100 ms intervals across a trial. The dependent measure was predicted as a function of scaled continuum step, and glottalization context (coded as in the categorization models), and the interaction of these two fixed effects in each time bin. Random effects were random intercepts for participant and random slopes for both fixed effects and their interaction. These models were run in *brms* as with models of the categorization data. The assessment of the effects will be in terms of when, in binned time, each has a robust effect on listeners' fixations.

As a point of comparison, a parallel moving window analysis from data in Steffman (2021a) (which was not reported in that paper) is presented here as well. Recall in that study, listeners heard a target word which was cued as prominent or not based on the pitch contour and amplitude of preceding material in a phrase. Steffman (2021a) found this phrasally cued prominence showed an overall delayed effect in processing (assessed by a Generalized Additive Mixed Model). The model was coded and implemented in the same way as the model of the Experiment 3 data. With this model we can make a direct comparison for the effect of continuum step and prominence cues (phrasal prominence versus glottalization) across experiments, and allowing us to examine if glottalization shows a different pattern from phrasal prominence in online processing.

In the second eyetracking analysis, eye movement data from Experiment 3 was analyzed by a Generalized Additive Mixed Model (GAMM) using the R packages *mgcv* (Wood, 2006) and *itsadug* (van Rij et al., 2016). GAMMs have recently been suggested to offer an appealing alternative to moving window analyses in that they allow for an encoding of the temporal contingency across time bins, and further allow for modeling non-linearity in the data (see Zahner et al., 2019 for a discussion of the advantages of GAMMs for eyetracking data). The data was sampled at 20 ms intervals for the GAMM analysis (as in Steffman, 2021a; Zahner et al., 2019). The GAMM model was fit with parametric terms for continuum step (scaled and centered), glottalization condition, and the interaction between these fixed effects. Parametric terms in the GAMM model are analogous to fixed effects in mixed effects models and capture if listeners' fixation preference in the analysis window as a whole varies as a function of the predictors. Smooth terms are additionally fit to model changes over time, and (potentially) non-linear patterns in the data. The model was fit to capture the interaction between continuum acoustics and time using a non-linear tensor-product interaction term, which allows us to examine how, over time, vowel acoustics mediate listeners' preference to fixate on a given target. There was an additional smooth term modeling the influence of glottalization condition over time, allowing us to examine the mediating influence of a preceding glottal stop. Random effects in the model were specified using the reference-difference smooth method described in Soskuthy (2021), with factors smooths for participant, and for participant by glottal stop condition (coded as an ordered factor). In both factor smooth terms, the *m* parameter was set to 1, following Baayen et al. (2018) and Soskuthy (2021). The numerical GAMM model output is included in the appendix, though the terms in the model as it was coded are generally not useful for intrepreting timecourse questions of interest here (Nixon et al., 2016; Zahner et al., 2019).

4.4 Timecourse predictions

Given the variables under consideration and the previous accounts of prosody and prominence in processing described in Section 1.2, we can operationalize some predictions in Experiment 3. First, a general expectation is that vowel-internal formant cues should exhibit a rapid influence in online processing as shown, for example, by Reinisch and Sjerps (2013). It takes approximately 200 milliseconds to program a saccadic eye movement, meaning that we expect a 200 ms lag between the time that a given stimulus dimension is used by listeners and the time it influences their looking behavior. Given this, we can predict to see an influence of vowel acoustics (modeled with the continuum variable) in online processing as early as 200 ms from the onset of the target vowel.

Taking this timing as a baseline for what constitutes a rapid effect, consider the timecourse predictions for vowel-initial glottalization. Following the prosodic analysis model, if a glottal stop is processed as strictly contributing to a prosodic parse of the signal which is integrated later in word recognition following Cho et al. (2007), it should show a later-stage effect in line with Kim et al. (2018b) and Mitterer et al. (2019). This predicts asynchrony between the influence of vowel formants and the glottal stop, with the effects of formants coming first. This asynchrony was indeed documented for phrasal prominence by Steffman (2021a), in the GAMM analysis presented in that paper, such that we expect to replicate it in the moving window analysis of that data presented here. A rapid effect of glottalization would suggest that it is not being processed in the manner outlined above.

In the present data and that from Steffman (2021a), the prominence manipulation only *preceded the target vowel in time*, and the target itself is acoustically the same across prominence-manipulating conditions.

4.5 Results

As shown in Figure 3, panel A, categorization results from Experiment 3 essentially replicated Experiment 1. Formant cues from the continuum exerted a reliable influence in categorization, (β = -2.61, 95%CrI = [-2.99, -2.26]; pd = 100), and we can see the categorization function is overall fairly well-anchored. The glottal stop effect from Experiment 1 was also replicated, with the presence of a preceding glottal stop increasing listeners' /ɛ/ responses (β = 2.49, 95%CrI = [1.99, 3.01]; pd = 100). An overall bias towards /ɛ/ is also evident in the tendency of listeners to categorize the target as /ɛ/, especially when it is preceded by a glottal stop.

Figure 3 panel B shows the raw eye movement data from the experiment, plotting eye movement trajectories as a function of continuum step and glottalization condition. The measure plotted on the y axis is listeners' preference to fixate on /ɛ/, computed as the proportion of looks to /ɛ/ minus looks to /æ/ in each 20 ms time bin. Here a value of zero indicates no preference, a positive value indicates a preference to fixate on /ɛ/ and a negative value indicates a preference to fixate on /æ/. We can see the effect of continuum step in the separation of lines based on coloration, with more /ɛ/-like continuum acoustics leading to a preference to fixate on /ɛ/. This separation, or fanning out, of trajectories appears to occur at roughly 200 ms from the onset of the vowel. The effect of vowel-initial glottalization is also evident in the separation we see based on line type: In line with the categorization data, we can see that a preceding glottal stop (dashed lines) facilitates looks to /ɛ/, an online effect corresponding to the categorization results we have seen thus far. We can also note that there is an /ɛ/-bias in eye movements, as also suggested by the categorization data, with steps 1- 4 showing a strong /ɛ/ preference. We can conclude that, qualitatively, it appears that both vowel-internal acoustic cues, and preceding glottalization, are jointly shaping listeners' perception of the target word. We now turn to the timing of these effect with respect to one another.

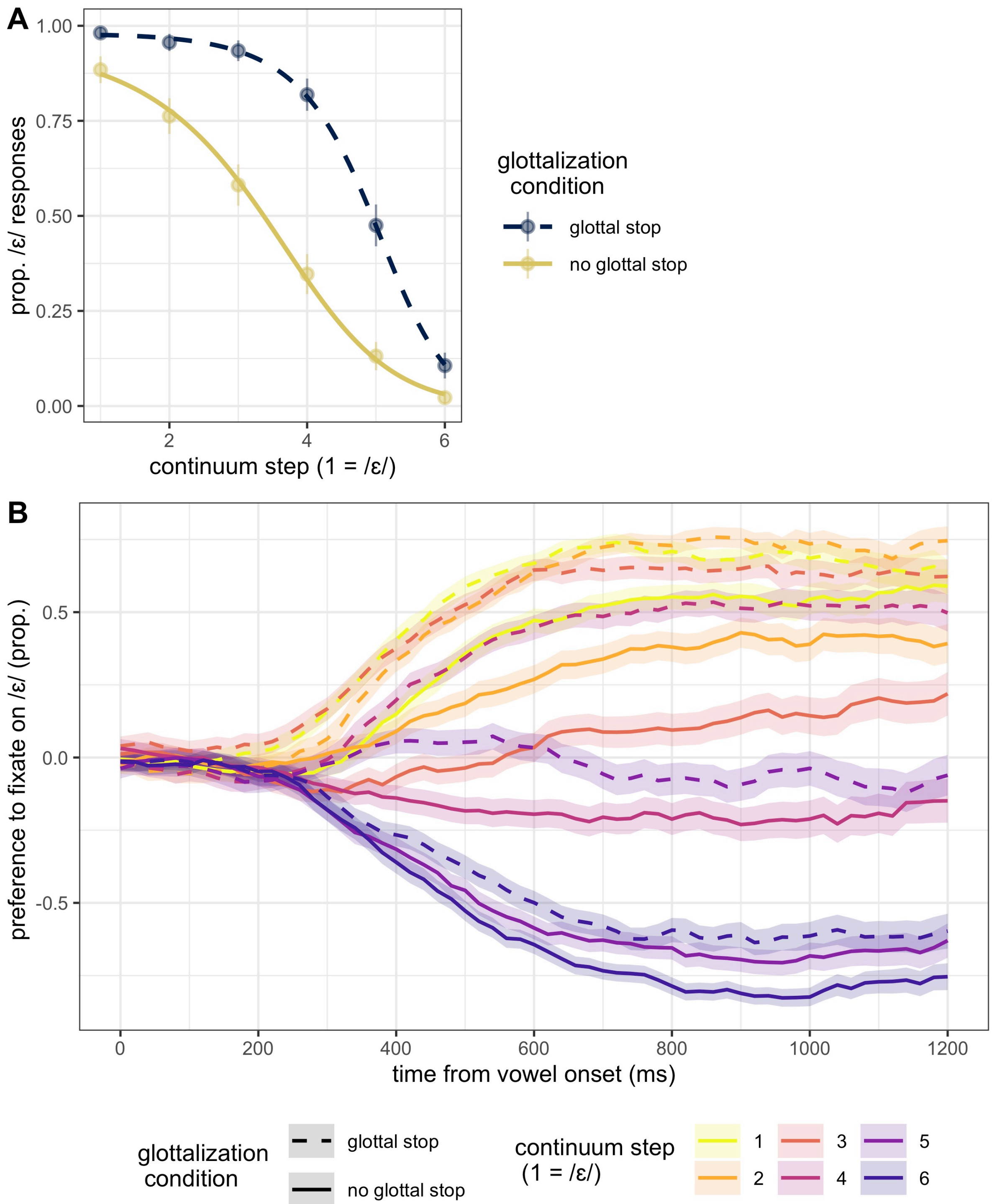

Figure 3: Categorization results in Experiment 3 (panel A), and eye movement data in Experiment 3 (panel B; see text). Error bars and ribbons show one SE, computed from the data.

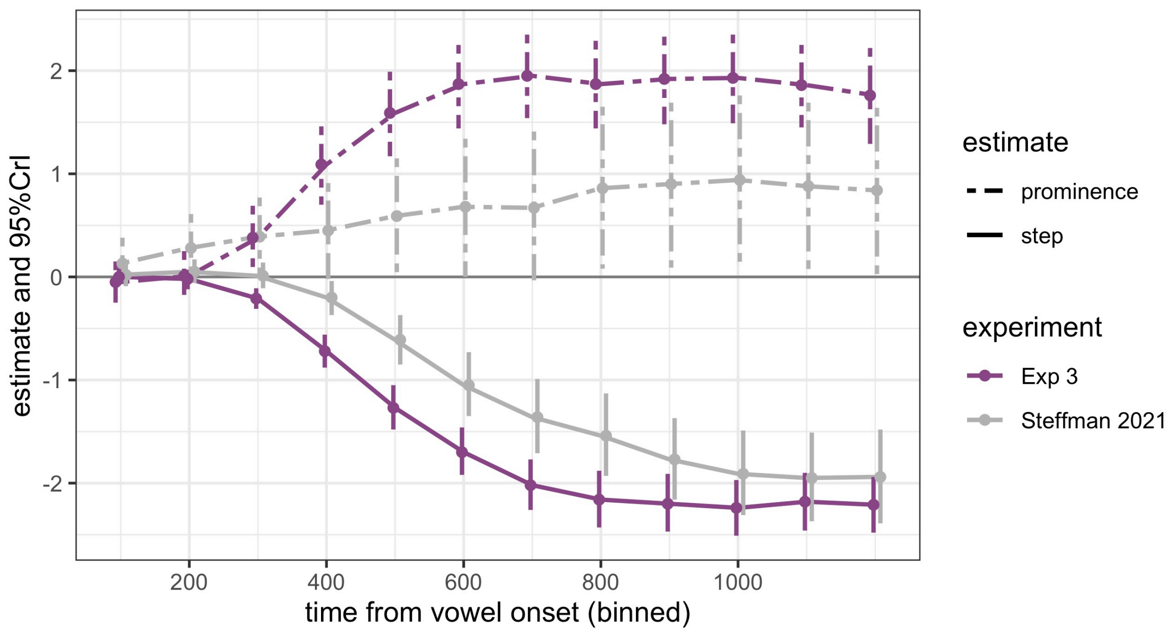

Figure 4: Model estimates for the effect of continuum step and prominence (glottalization) in the moving window analysis for Experiment 3, with estimates from the same analysis for data from Steffman (2021a) for comparison. Each point is located at the end of a time bin, e.g., 200 indicates 100-200 ms.

4.5.1 Moving window analysis

In the moving window analysis, estimate for each effect from the model is given along with 95% CrI, each of which are plotted over time, which is presented in 100 ms time bins, shown in Figure 4. The full model summaries which produced the estimates plotted here are contained in the open access repository.

First consider just the data from Experiment 3. The timing of an effect can be taken to be reliable is the time bin in which 95%CrI for the effect *exclude* the value of 0. A reliable effect of continuum step in Experiment 3 evident in the 200-300 ms time bin (note the estimates are arbitrarily negative because of the way in which the variables were coded, i.e. decreases in the log-transformed /ɛ/-preference as a function of increasing values of continuum step). This effect is early and is consistent with previous work showing a rapid use of vowel formants in processing vowel information (Reinisch and Sjerps, 2013). Next, consider the timing of the step effect in relation to the glottal stop effect (labeled as the prominence effect for Experiment 3). This effect also becomes credibly different from zero at the same time as the effect of continuum step (200-300 from target onset). The influence of continuum step and the glottal stop are thus simultaneous.

This simultaneous effect can be compared to the timing of the effects of vowel acoustics and phrasal prominence from Steffman (2021a), also plotted in Figure 4. The effect of continuum step is reliable 300-400 ms from the onset of the target vowel, one time bin later than the effect of continuum step in Experiment 3. The effect of the phrasal prominence manipulation in Steffman (2021a) is smaller in size compared to Experiment 3, and does not show a consistent divergence from 0 until the 700-800ms time bin and onward (though there is a transitory and smaller credible effect at 400-500 ms). This lines up with the GAMM analysis presented in Steffman (2021a), which showed subtle effects of phrasal prominence early in time, with larger and more robust effects only apparent later in the analysis window. Importantly, the robust effect is clearly asynchronous with the effect of vowel acoustics in that experiment, differentiating it from the synchronous influence of a glottal stop, and vowel formants, in online processing.

The moving window analysis thus allows us to conclude that the influence of vowel initial glottalization is rapid, and synchronous with the influence of vowel formants. It is further different in its timing from the effect of phrasal prominence, a point that is returned to in section 5 below.

4.5.2 GAMM modeling

Given the simultaneous effect of continuum step and vowel acoustics, the GAMM modeling analysis focused on the relationship between glottalization and formants in jointly shaping listeners' processing of the target word. First, the parametric terms in the GAMM model confirm an influence of vowel formants and glottalization in the analysis window as a whole ($p < 0.001$ for both), as would be expected given the moving window results.

To assess the relationship between continuum step, glottal stop condition, and time, three dimensional topographic surface plots are presented in Figure 5. These plots show the model fit, representing the effect of continuum step (as a continuous variable on the y axis) over time (on the x axis). The dependent variable (listeners' Elog-transformed preference to fixate on the /ɛ/ target) is represented on a gradient color scale. The two panels represent model fits based on glottalization condition, the top panel being when the target is preceded by a glottal stop. A value of zero (in the middle of the color scale) indicates no preference, while a positive value (closer to yellow on the color scale) indicates a preference for the /ɛ/ target. A negative value (closer to purple on the color scale) represents a preference for /æ/. Shading on the surface shows locations where listeners' preference is not significantly different than zero, i.e.

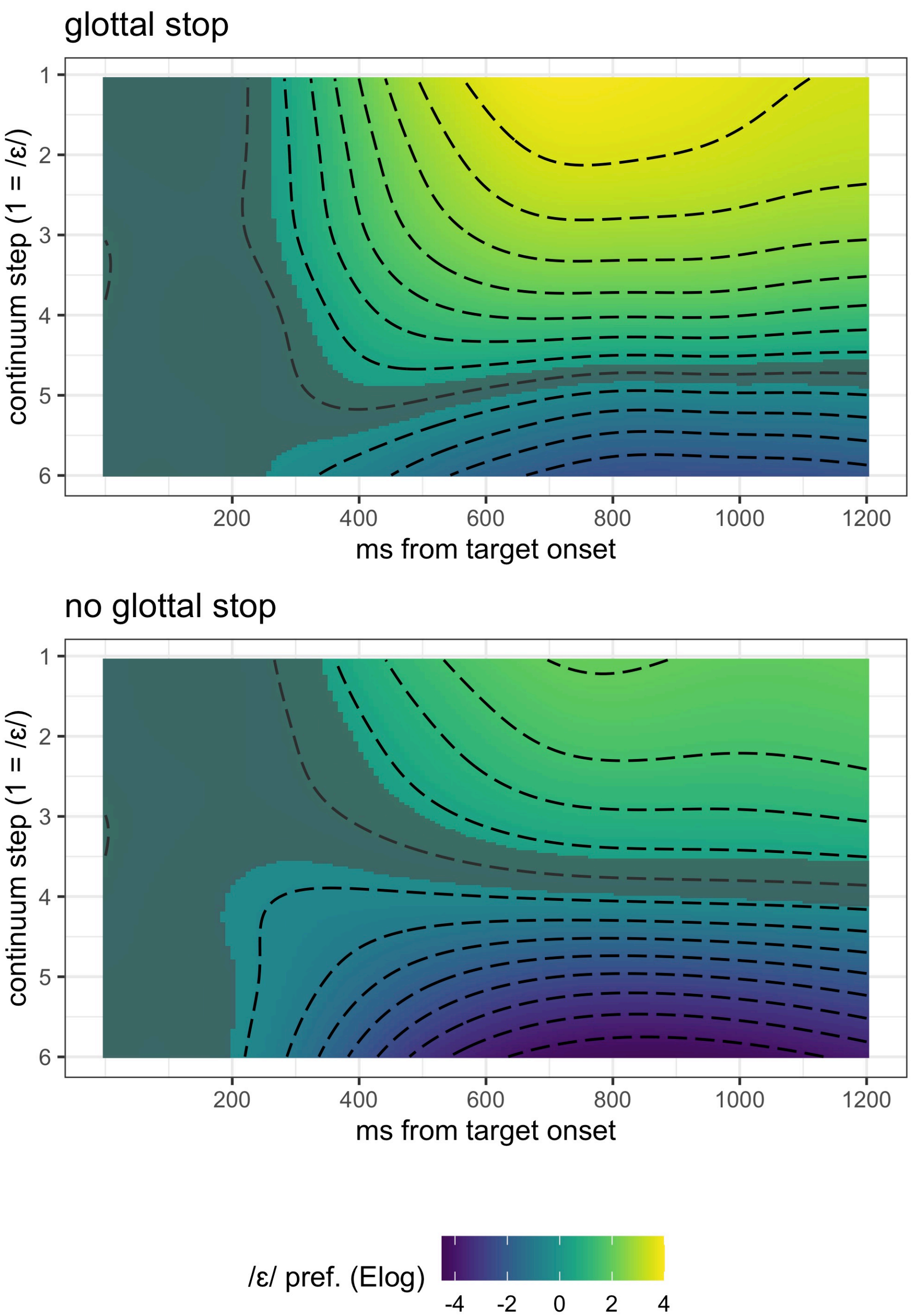

Figure 5: Surface plots showing the GAMM model fit in Experiment 3, with continuum step on the y axis, time on the x axis, and listeners' log-transformed fixation preference indexed by coloration. Gray shading indicates places on the surface where listeners have no preference for either target.

when 95% CI from the model estimate include the value of zero. Note that listeners do not show a preference early in the analysis window (with shading on all of the surface prior to approximately 200 ms).

As time progresses, listeners develop graded preferences based on continuum step. At the end of the analysis window, there is a range of preferences: a stronger /ɛ/ preference at step 1 on the continuum, and a stronger /æ/ preference at step 6. Note too that some portion in the middle region of the continuum never attains a significant preference in either panel. That is, the model finds that the ambiguous region of the continuum remains ambiguous even at the end of the analysis window. This is shown by the shaded area persisting until end of the analysis window. With this in mind, we now can assess the impact of a glottal stop on listeners' use of the continuum over time. The effect of the glottal stop is evident in observing (1) the coloration of each panel A and B, and (2) the shape and position of the shaded area showing points on the surface for which listeners did not have a preference for either target. In terms of coloration, note the color scale used in both panels is shared by them: the same color on each panel would reflect the same degree of /ɛ/ preference. We can see that each panel overall occupies different color spaces, with the glottal stop condition showing a stronger /ɛ/ preference (more yellow on the plot), and the no glottal stop condition showing a stronger /æ/ preference (more purple on the plot). In other words, acoustically identical continuum steps are perceived as more like one target or the other, as function of glottalization. These differences are evident as early as listeners show *any* preference: as soon as the shading on the surfaces disappears.[4]

Additionally, the surface plots show that glottal stop 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 glottal stop condition, starting early (roughly 200 ms from target onset) and persisting throughout the analysis window. This is another piece of evidence that the glottal stop is shaping listeners' perception of formant cues directly. Inspection of the surface plots therefore supports a difference in early formant processing across conditions, with differences across conditions evident at the earliest moments expected, and early modulation of which vowel acoustics are ambiguous to listeners.

4.6 Discussion

The timecourse data for Experiment 3 can be summarized as showing a rapid influence of vowel-initial glottalization in vowel perception, in line with sonority expansion effects on vowel formants. This influence was rapid in the sense that it was synchronous with the immediate use of vowel formants as determined by the moving window analysis, and impacted fixations as soon as listeners showed a preference for any target, as determined by the GAMM analysis. It also preceded the effect of phrasal prominence in time, as shown by comparing the timing of the effect to data from (Steffman, 2021a). This comparison shows that different sorts of prominence lending information can impact processing at different times, and effects can vary both magnitude and timing. Implications for this rapid effect of vowel-initial glottalization and difference in effect timing are discussed in more detail below.

5 General Discussion

The present study set out to examine if listeners are impacted by the presence of vowel-initial glottalization in their perception of vowel contrasts. In Experiment 1, it was seen that the production of a sustained glottal stop preceding a vowel led to listeners mapping vowel acoustics to a lower/backer realization of the vowel, in line with the ways in which the relevant vowel acoustics are modulated by prominence. Experiment 2 showed that these effects are also evident when glottalization is cued by dipping pitch and intensity, and that stronger glottalization cues lead to larger perceptual shifts. Intermediate steps on the glottalization continuum led to intermediate shifts in categorization, suggesting that stronger vowel-initial glottalization cued a stronger percept of prominence. Experiment 3 replicated the effects of a full glottal stop seen in Experiment 1, in a visual-world eyetracking paradigm which compared the timecourse of the influence of a preceding glottal stop to that of vowel-internal formant values. Both of these influences were simultaneous, with a vowel-initial glottal stop immediately impacting perception and modulating how formant cues are used at the earliest moments in processing.

Let us first consider these results as they relate to the hypothesized prominence-marking function of word-initial glottalization in American English in the speech production literature. In line with prominence cued phrasally in Steffman (2021a), the presence of glottalization preceding a vowel led to listeners' expectation of a more prominent (in this case, sonorous) variant of that vowel being produced. This data thus supports the proposal that glottalization cues prominence to listeners, in line with its implementation as a prominence marker in production. This interpretation more generally accords with Mitterer et al. (2021a,b) in that glottalization is an important prosodic cue which is recruited in perception by listeners.

It is worth noting here that across all conditions in the present experiments the target word was pitch accented, such that the prominence effects seen here suggest different levels of perceptual prominence within pitch accented words, and fine-grained variation in prominence perception as shown in Experiment 2. This approach fits with recent conceptualizations of prominence in terms of both phonological and phonetic features (Baumann and Cangemi, 2020). If we consider "pitch-accented" to be a phonological specification of prominence category, these results speak to the importance of considering within-category variation in perceived prominence as meaningfully impacting the perception of segmental material, in line too with Dilley et al. (1996) showing that pitch accentuation predicts vowel-initial glottalization but does not require it.

More broadly, this result suggests that future research will benefit from considering other patterns of prominence strengthening as relevant in segmental perception. For example, consider the lengthening of VOT in voiceless stops which is observed in prominent syllables (Cole et al., 2007; Kim et al., 2018a). Given the present results we can predict that prominence-signaling lengthening of VOT may impact perception of the following vowel. If found, this would further indicate the importance of fine-grained prominence-strengthening cues in segmental perception. A key takeaway from these results is accordingly the view that prosody should be considered not only in terms of suprasegmental parameters, nor strictly abstract structural terms (phrase boundaries, pitch accents) but should be viewed holistically and as encoded in fine-grained detail and modulation of cues such as VOT and formant structure as well.

The eyetracking data further enrich our understanding of the interplay between prosodic and segmental/lexical processing. As noted in Section 1.2, examination of prosodic influences in segmental processing have focused on manipulations of more global phrasal prosodic context, with data supporting a delayed influence of prosodic structure, overall consistent with a post-lexical model of prosodic effects. Such an account of the present data predicts a two-stage influence of segment-internal cues to a contrast and prosodic context, with segmental cues preceding prosodic context in the timecourse of their influence. The data in Experiment 3 are not consistent with this account, with *simultaneous* effects of formants and a preceding glottal stop in online processing. This effect is clearly different from the effect of phrasal prominence in being robust early in processing, as described in section 4.6. This asymmetry suggests more generally that prominence effects can unfold at multiple time scales, consistent with a local to global perception of prominence that may feed forward into a more abstract representation of prosodic structure, as described for the prosodic analysis model by Cho et al. (2007).

As previous work on phrasal prosodic boundaries in processing show clear support for only a delayed influence of prosodic boundary information in the perception of segmental material (Kim et al., 2018b; Mitterer et al., 2019), the present data suggest the field will benefit from considering that prominence information and prosodic boundary information may enter differently into processing. One possible view of the asymmetrical role of these prosodic dimensions is that prosodic boundary information is necessarily structural: the listener must determine the presence a boundary based on phonetic cues, broader phonological context, word boundary information, and syntactic information. Inferences about these levels of representation can be presumed to take place in parallel, and with the consideration of multiple hypotheses, framed recently through the lens of Bayesian inference by McQueen and Dilley (2020). Prominence could also be described as structural in the sense that in American English (among other languages) it is determined based on metrical structure and phrasing (e.g., the most prominent pitch accent, the nuclear accent is the last one in an intonational phrase). However prominence should also clearly be viewed at a more fine-grained level as the property of perceptually "standing out" from context (Baumann and Cangemi, 2020), something we could consider to be "phonetic prominence". In this sense the present study shows the importance of considering phonetic prominence, signaled by language-specific cues such as vowel-initial glottalization, as being incorporated by listeners in their perception of segments. The determination of a given unit's prominence in this sense therefore needn't be determined by a more global prosodic parse, but instead may be computed by the listener on a syllable by syllable (or even perhaps in some cases segment by segment) basis. Prominence in this regard can be useful in determining if a segment has undergone prominence strengthening effects: in essence, in reconciling the extent to which a segment is perceptually prominent, with its acoustic structure to determine how it should map to a phonemic category. This could be re-framed as the listener computing "is this segment phonetically strengthened based on prominence?" This view implicates prominence at both sub-lexical and higher levels, in multiple stages of processing. Steffman (2020) accordingly argued that prosodic prominence merits consideration both in terms of structural (or, phonological) prosodic prominence, and its phonetic encoding, and proposed the *Multistage Assessment of Prominence in Processing* model, a two-stage model, which is supported by the data here showing that different sorts of prominence-lending cues seems to be integrated in perception at different time scales.

Further tests for this sort of distinction between localized/phonetic and global/structural prominence cues could take the form of examining the extent to which each can be modulated by task factors. Certain early effects in processing are assumed to be relatively immune to task effects and cognitive load as shown by, e.g., Bosker et al. (2017). More global prosodic factors have recently been shown to be influenced by task and stimulus presentation factors (Steffman, 2019, 2021b). For example, Steffman (2021b) found that rhythmic effects in the perception of segmental cues are disrupted when stimuli vary in speech rate, while speech rate effects (typically assumed to result from low-level auditory processing) are robust to rhythmic variation and occur consistently. To the extent that the effects of vowel-initial glottalization seen here reflect early sublexical processing we might expect them to be robust to these sorts of task effects whereas global prominence effects may be more fragile.

In this vein, one outstanding question is the extent to which localized prominence strengthening effects are related to more general auditory processing. Though glottalization as prominence strengthening is certainly implemented in a language-specific fashion by speakers, it has the effect of making the following vowel acoustically prominent in a more general way (i.e. a vowel preceded by glottalization is rendered louder than, and perceptually more separated from, preceding material) which boosts auditory processing (Delgutte, 1980; Delgutte and Kiang, 1984). Pulling apart the role of language-specific phonetic knowledge and language-general prominence perception may be difficult as phonetic strengthening patterns tend to serve the function of making the strengthened segment more prominent perceptually (though Steffman, 2020 shows that the effects of prominence on vowel perception are specific to the vowel contrast in question). Some indirect evidence for a language-specific interpretation of glottalization cues comes from comparing the early time course of the effect seen here to the delayed influence documented in Mitterer et al. (2019), where a delayed effect is consistent with higher level prosodic analysis. This suggests that the processing of glottalization for American English listeners is different from its processing in Maltese. Carefully controlled cross-linguistic experiments may be useful as a further test of language-general versus language-specific effects going forwards, particularly across languages in which glottalization patterns and functions differently.

Future work will also benefit from examining the ways in which both early and later effects of prosodic organization are weighted against other components of linguistic structure. For example, effects of segmental co-occurrence probability (often computed as biphone probability) are typically assumed to operate at a sublexical level of processing (Norris et al., 2000; Pitt and McQueen, 1998), while effects related to word-hood and neighborhood density entail contact with the lexicon and a hypothesized delay in processing (cf. Newman et al., 1997; Kingston et al., 2016). Comparing these effects in their timecourse to those that we see for prominence-related effects and examining the extent to which they interact and compete in determining the listener's interpretation of speech will be a useful test going forwards. For example, if the effects documented here represent sublexical processing we should expect similar timing between them and effects of biphone probability, and a joint influence of these factors in determining listeners' perception.

In sum, relating the present results to other phonetic strengthening patterns, other languages, and other aspects of linguistic organization will help situate these findings with our understanding of the detailed interplay between segmental and prosodic processing in speech comprehension.

Acknowledgments

Many thanks are due Adam Royer for recording stimuli for the experiments, to Danielle Frederickson, Qingxia Guo and Bryan Gonzalez for help with data collection, and to Sun-Ah Jun, Pat Keating, Megha Sundara and Taehong Cho for valuable feedback and discussion.

Notes

[1]Spectral contrast refers to the perception of frequency regions in the spectrum (here, formants) relative to contextual spectral information (Stilp, 2020; Holt et al., 2000). Contrast effects diminish in strength as there is increased distance between context and target (Holt, 2005; Stilp, 2018). Contrast effects here will thus be strongest in the no glottal stop condition, where no glottal stop temporally separates the precursor and the target. In the present stimuli, the precursor vowel generally has higher F1 and lower F2 than the formant values on the continuum. Thus, F1 will be perceived as relatively low and F2 will be perceived as relatively high in the target (more like /ɛ/) as a function of spectral contrast with the precursor. This predicts that the target is more likely to be perceived as /ɛ/ in the no glottal stop condition, where contrast effects should be strongest. This is the opposite of the prediction based on glottalization as a prominence cue, described in Section 3. Note that if the precursor had consistently lower F1 and higher F2 than the target continuum, spectral contrast would be a confound.

[2]The 0 mean of the prior for the intercept encodes a expectation of equal odds of "ebb" versus "ab" responses at the center of the continuum, as the continuum variable is centered and scaled. The 0 mean of

the prior for the fixed effects encodes a prior expectation a change of 0 in log odds as a function of either fixed effect (i.e., no prior expectation of an effect). The standard deviation of 1.5 (in log-odds) encodes a wide window of uncertainly around these values, which is essentially flat in log-odds space (McElreath, 2020), i.e., high uncertainty about what the effects will be in both magnitude and directionality. Such priors thus provide some information to the model about the intercept but are very weakly informative, allowing for the data to "speak for itself". This is appropriate for hypothesis testing of the sort carried out here where there is not any prior expectation about the data , see e.g., McElreath, 2020.

[3]The transformation is the following, where n is the total number of samples in a given time bin and y is the number of samples for a given interest area:

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

[4]We can also note tangentially that slightly more of the surface overall is shaded when there is no glottal stop (32%), as compared to when there is a glottal stop (29%). This is consistent with the idea that the glottal stop facilitates recognition of the target vowel, allowing listeners to develop as fixation preference sooner overall, as compared to when no glottal stop precedes the target.

Appendix

Table 1: Model outputs for categorization results

	Estimate	Est. Error	L-95% CI	U-95%CI	pd
Experiment 1					
	Estimate	Est. Error	L-95% CI	U-95%CI	pd
intercept	1.15	0.15	0.83	1.45	100
glottal stop	1.69	0.22	1.26	2.12	100
continuum	-3.29	0.17	-3.62	-2.97	100
glottal stop:continuum	-0.75	0.19	-1.13	-0.39	100
Experiment 2					
	Estimate	Est. Error	L-95% CI	U-95%CI	pd
intercept	0.75	0.12	0.87	1.50	100
glottalization (scaled)	0.38	0.05	0.29	0.49	100
continuum	-2.95	0.16	-3.27	-2.64	100
glottalization:continuum	-0.24	0.07	-0.40	-0.11	100
Experiment 3					
	Estimate	Est. Error	L-95% CI	U-95%CI	pd
intercept	0.95	0.16	0.66	1.26	100
glottal stop	2.49	0.26	1.99	3.01	100
continuum	-2.61	0.18	-2.99	-2.26	100
glottal stop:continuum	-0.43	0.17	-0.78	-0.11	100

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

Parametric terms	Estimate	Est. Error	t-value	p-value
intercept	0.86	0.08	11.36	< 0.001
continuum	-0.84	0.04	-18.8	< 0.001
glottal stop	-1.32	0.13	-10.59	< 0.001
glottal stop:continuum	-0.06	0.06	-1.09	0.31
Smooth terms	edf	ref df	F-value	p-value
te(time, continuum)	26.74	29.98	85.35	< 0.001
te(time, continuum; condition = glottal stop)	8.20	8.56	12.38	< 0.001
s(time, continuum; condition = no glottal stop)	2.21	2.52	1.42	0.27
s(time, participant)	244.68	359.00	2.69	< 0.001
s(time, participant; condition)	207.19	359.00	1.61	< 0.001

References

Baayen, R. H., van Rij, J., de Cat, C., and 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*, pages 49–69. Springer. doi: https://doi.org/10.48550/arXiv.1601.02043.

Barr, D. J. (2008). Analyzing 'visual world' eyetracking data using multilevel logistic regression. *Journal of Memory and Language*, 59(4):457–474. doi: http://dx.doi.org/10.1016/j.jml.2007.09.002.

Baumann, S. and Cangemi, F. (2020). Integrating phonetics and phonology in the study of linguistic prominence. *Journal of Phonetics*, 81:100993. doi: https://doi.org/10.1016/j.wocn.2020.100993.

Beckman, M. E., Edwards, J., and Fletcher, J. (1992). *Prosodic structure and tempo in a sonority model of articulatory dynamics*, pages 68–89. Papers in Laboratory Phonology. Cambridge University Press. doi: https://doi.org/10.1017/CBO9780511519918.004.

Boersma, P. and Weenink, D. (2020). Praat: doing phonetics by computer (version 6.1.09).

Bosker, H. R., Reinisch, E., and 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. doi: https://doi.org/10.1016/j.jml.2016.12.002.

Brand, S. and Ernestus, M. (2018). Listeners' processing of a given reduced word pronunciation variant directly reflects their exposure to this variant: Evidence from native listeners and learners of french. *Quarterly Journal of Experimental Psychology*, 71(5):1240–1259. doi: https://doi.org/10.1080/17470218.2017.1313282.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1):1–28. doi: https://doi.org/10.18637/jss.v080.i01.

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. doi: https://doi.org/10.1121/1.1861893.

Cho, T., McQueen, J. M., and 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. doi: https://doi.org/10.1016/j.wocn.2006.03.003.

Chong, J. and Garellek, M. (2018). Online perception of glottalized coda stops in American English. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*. doi: ttps://doi.org/10.5334/labphon.70.

Christophe, A., Peperkamp, S., Pallier, C., Block, E., and Mehler, J. (2004). Phonological phrase boundaries constrain lexical access I. Adult data. *Journal of Memory and Language*, 51(4):523–547. doi: https://doi.org/10.1016/j.jml.2004.07.001.

Cole, J., Kim, H., Choi, H., and Hasegawa-Johnson, M. (2007). Prosodic effects on acoustic cues to stop voicing and place of articulation: Evidence from Radio News speech. *Journal of Phonetics*, 35(2):180–209. doi: https://doi.org/10.1016/j.wocn.2006.03.004.

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. doi: https://doi.org/10.1121/1.412275.

Delgutte, B. (1980). Representation of speech-like sounds in the discharge patterns of auditory-nerve fibers. *The Journal of the Acoustical Society of America*, 68(3):843–857. doi: https://doi.org/10.1121/1.384824.

Delgutte, B. and Kiang, N. Y. (1984). Speech coding in the auditory nerve: I. vowel-like sounds. *The Journal of the Acoustical Society of America*, 75(3):866–878. doi: https://doi.org/10.1121/1.390599.

Dilley, L., Shattuck-Hufnagel, S., and Ostendorf, M. (1996). Glottalization of word-initial vowels as a function of prosodic structure. *Journal of Phonetics*, 24(4):423–444. doi: https://doi.org/10.1006/jpho.1996.0023.

Erickson, D. (2002). Articulation of extreme formant patterns for emphasized vowels. *Phonetica*, 59(2-3):134–149. doi: https://doi.org/10.1159/000066067.

Garellek, M. (2013). *Production and perception of glottal stops*. PhD thesis, University of California, Los Angeles.

Garellek, M. (2014). Voice quality strengthening and glottalization. *Journal of Phonetics*, 45:106–113. doi: https://doi.org/10.1016/j.wocn.2014.04.001.

Garellek, M. and Keating, P. (2011). The acoustic consequences of phonation and tone interactions in Jalapa Mazatec. *Journal of the International Phonetic Association*, pages 185–205. doi: https://doi.org/10.1017/S0025100311000193.

Garellek, M. and White, J. (2015). Phonetics of Tongan stress. *Journal of the International Phonetic Association*, 45(01):13–34. doi: https://doi.org/10.1017/S0025100314000206.

Gerfen, C. and Baker, K. (2005). The production and perception of laryngealized vowels in Coatzospan Mixtec. *Journal of Phonetics*, 33(3):311–334. doi: https://doi.org/10.1016/j.wocn.2004.11.002.

Gordon, M. and Ladefoged, P. (2001). Phonation types: a cross-linguistic overview. *Journal of Phonetics*, 29(4):383–406. doi: https://doi.org/10.1006/jpho.2001.0147.

Holt, L. L. (2005). Temporally nonadjacent nonlinguistic sounds affect speech categorization. *Psychological Science*, 16(4):305–312. doi: https://doi.org/10.1111/j.0956-7976.2005.01532.x.

Holt, L. L., Lotto, A. J., and Kluender, K. R. (2000). Neighboring spectral content influences vowel identification. *The Journal of the Acoustical Society of America*, 108(2):710–722. doi: https://doi.org/10.1121/1.429604.

Huffman, M. K. (2005). Segmental and prosodic effects on coda glottalization. *Journal of Phonetics*, 33(3):335–362. doi: https://doi.org/10.1016/j.wocn.2005.02.004.

Jongenburger, W. and van Heuven, V. J. (1991). The distribution of (word initial) glottal stop in Dutch. *Linguistics in the Netherlands*, 8(1):101–110. doi: https://doi.org/10.1075/avt.8.13jon.

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

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

Kim, S. and Cho, T. (2013). Prosodic boundary information modulates phonetic categorization. *The Journal of the Acoustical Society of America*, 134(1):EL19–EL25. doi: https://doi.org/10.1121/1.4807431.

Kim, S., Kim, J., and Cho, T. (2018a). Prosodic-structural modulation of stop voicing contrast along the VOT continuum in trochaic and iambic words in American English. *Journal of Phonetics*, 71:65–80. doi: https://doi.org/10.1016/j.wocn.2018.07.004.

Kim, S., Mitterer, H., and Cho, T. (2018b). A time course of prosodic modulation in phonological inferencing: The case of Korean post-obstruent tensing. *PloS one*, 13(8). doi: https://doi.org/10.1371/journal.pone.0202912.

Kingston, J., Levy, J., Rysling, A., and Staub, A. (2016). Eye movement evidence for an immediate Ganong effect. *Journal of Experimental Psychology: Human Perception and Performance*, 42(12):1969. doi: https://doi.org/10.1037/xhp0000269.

Kreiman, J. and Sidtis, D. (2011). *Foundations of voice studies: An interdisciplinary approach to voice production and perception*. John Wiley & Sons. doi: https://doi.org/10.1002/9781444395068.

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

Maddieson, I. and Precoda, K. (1989). Updating UPSID. *The Journal of the Acoustical Society of America*, 86(S1):S19–S19. doi: https://doi.org/10.1121/1.2027403.

Makowski, D., Ben-Shachar, M. S., and Lüdecke, D. (2019). bayestestr: Describing effects and their uncertainty, existence and significance within the bayesian framework. *Journal of Open Source Software*, 4(40):1541. doi: https://doi.org/10.21105/joss.01541.

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC.

McQueen, J. M. and Dilley, L. (2020). Prosody and spoken-word recognition. In *The Oxford handbook of language prosody*, pages 509–521. Oxford University Press. doi: https://doi.org/10.1093/oxfordhb/9780198832232.013.33.

Mendelsohn, A. H. and Zhang, Z. (2011). Phonation threshold pressure and onset frequency in a two-layer physical model of the vocal folds. *The Journal of the Acoustical Society of America*, 130(5):2961–2968. doi: https://doi.org/10.1121/1.3644913.

Mitterer, H., Cho, T., and Kim, S. (2016). How does prosody influence speech categorization? *Journal of Phonetics*, 54:68–79. doi: https://doi.org/10.1016/j.wocn.2015.09.002.

Mitterer, H., Kim, S., and Cho, T. (2019). The glottal stop between segmental and suprasegmental processing: The case of Maltese. *Journal of Memory and Language*, 108:104034. doi: https://doi.org/10.1016/j.jml.2019.104034.

Mitterer, H., Kim, S., and Cho, T. (2021a). Glottal stops do not constrain lexical access as do oral stops. *PloS one*, 16(11):e0259573. doi: https://doi.org/10.1371/journal.pone.0259573.

Mitterer, H., Kim, S., and Cho, T. (2021b). The role of segmental information in syntactic processing through the syntax–prosody interface. *Language and Speech*, 64(4):962–979. doi: https://doi.org/10.1177/0023830920974401.

Mitterer, H. and 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. doi: https://doi.org/10.1016/j.jml.2013.07.002.

Mo, Y., Cole, J., and Hasegawa-Johnson, M. (2009). Prosodic effects on vowel production: evidence from formant structure. In *Proceedings of INTERSPEECH*, pages 2535–2538.

Moulines, E. and Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5-6):453–467. doi: https://doi.org/10.1016/0167-6393(90)90021-Z.

Newman, R. S., Sawusch, J. R., and Luce, P. A. (1997). Lexical neighborhood effects in phonetic processing. *Journal of Experimental Psychology: Human Perception and Performance*, 23(3):873. doi: https://doi.org/10.1037/0096-1523.23.3.873.

Nixon, J. S., van Rij, J., Mok, P., Baayen, R. H., and 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. doi: https://doi.org/10.1016/j.jml.2016.03.005.

Norris, D., McQueen, J. M., and Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences*, 23:299–325. doi: https://doi.org/10.1017/S0140525X00003241.

Pierrehumbert, J. B. and Frisch, S. (1997). Synthesizing allophonic glottalization. In *Progress in Speech Synthesis*, pages 9–26. Springer. doi: https://doi.org/10.1007/978-1-4612-1894-4_2.

Pierrehumbert, J. B. and Talkin, D. (1992). *Lenition of /h/ and glottal stop*, pages 90–127. Papers in Laboratory Phonology. Cambridge University Press. doi: https://doi.org/10.1017/CBO9780511519918.005.

Pitt, M. A. (2009). How are pronunciation variants of spoken words recognized? A test of generalization to newly learned words. *Journal of Memory and Language*, 61(1):19–36. doi: https://doi.org/10.1016/j.jml.2009.02.005.

Pitt, M. A. and McQueen, J. M. (1998). Is compensation for coarticulation mediated by the lexicon? *Journal of Memory and Language*, 39(3):347–370. doi: https://doi.org/10.1006/jmla.1998.2571.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reinisch, E. and 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. doi: https://doi.org/10.1016/j.wocn.2013.01.002.

RStudio Team (2021). *RStudio: Integrated Development Environment for R*. RStudio, PBC., Boston, MA.

Silverman, K. and Pierrehumbert, J. (1990). The timing of prenuclear high accents in English. In Beckman, M. E. and Kingston, J., editors, *Papers in Laboratory Phonology*, Papers in Laboratory Phonology, pages 72–106. doi: https://doi.org/10.1121/1.2024693.

Soskuthy, M. (2021). Evaluating generalised additive mixed modelling strategies for dynamic speech analysis. *Journal of Phonetics*, 84:101017. doi: https://doi.org/10.1016/j.wocn.2020.101017.

Steffman, J. (2019). Phrase-final lengthening modulates listeners' perception of vowel duration as a cue to coda stop voicing. *The Journal of the Acoustical Society of America*, 145(6):EL560–EL566. doi: https://doi.org/10.1121/1.5111772.

Steffman, J. (2020). *Prosodic Prominence in Vowel Perception and Spoken Language Processing*. PhD thesis, University of California, Los Angeles.

Steffman, J. (2021a). Prosodic prominence effects in the processing of spectral cues. *Language, Cognition and Neuroscience*, 36(5):586–611. doi: https://doi.org/10.1080/23273798.2020.1862259.

Steffman, J. (2021b). Rhythmic and speech rate effects in the perception of durational cues. *Attention, Perception, & Psychophysics*, 83(8):3162–3182. doi: https://doi.org/10.3758/s13414-021-02334-w.

Stilp, C. (2018). Short-term, not long-term, average spectra of preceding sentences bias consonant categorization. *The Journal of the Acoustical Society of America*, 144(3):1797–1797. doi: https://doi.org/10.1121/1.5067927.

Stilp, C. (2020). Acoustic context effects in speech perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, 11(1):e1517. doi: https://doi.org/10.1002/wcs.1517.

Tehrani, H. (2020). Appsobabble: Online applications platform.

Traunmüller, H. (1990). Analytical expressions for the tonotopic sensory scale. *The Journal of the Acoustical Society of America*, 88(1):97–100. doi: https://doi.org/10.1121/1.399849.

van Rij, J., Wieling, M., Baayen, R., and van Rijn, H. (2016). itsadug: Interpreting time series and autocorrelated data using GAMMs [R package].

Winn, M. (2016). Vowel formant continua from modified natural speech (Praat script). Version 38.

Wood, S. N. (2006). *Generalized Additive Models: an Introduction with R*. Chapman and Hall/CRC. doi: https://doi.org/10.1201/9781420010404.

Zahner, K., Kutscheid, S., and Braun, B. (2019). Alignment of f0 peak in different pitch accent types affects perception of metrical stress. *Journal of Phonetics*, 74:75–95. doi: https://doi.org/10.1016/j.wocn.2019.02.004.

Zhang, Z. (2011). Restraining mechanisms in regulating glottal closure during phonation. *The Journal of the Acoustical Society of America*, 130(6):4010–4019. doi: https://doi.org/10.1121/1.3658477.