Prosodic phrasing mediates listeners' perception of temporal cues: Evidence from the Korean Accentual Phrase

Jeremy Steffman[a], Sahyang Kim[b], Taehong Cho[c], Sun-Ah Jun[d]

[a]*Northwestern University*
[b]*Hongik University*
[c]*Hanyang University*
[d]*University of California, Los Angeles*

Abstract

In two experiments we examine how listeners make reference to prosodic phrasing in their perception of temporally cued segmental contrasts. We test how the prosodic-structurally conditioned modulation of segmental cues (in domain-initial strengthening) translates into speech perception. We adopt the test case of stop contrasts in Seoul Korean (aspirated versus fortis), which are cued by vowel duration and voice onset time (VOT). The phrasing manipulation was carried out at the level of the Accentual Phrase (AP), a small phrase that is marked by intonational features. The AP was chosen because it was possible to create two prosodic phrasing contexts (AP-initial versus AP-medial) by manipulating only f0 before the target segment with the duration of contextual segments unchanged, controlling for temporal context effects. In Experiment 1, listeners shift their perception of a VOT continuum based on phrasing, in line with the domain-initial strengthening pattern of post-stop vowel lengthening, where AP-initial post-fortis vowels are lengthened. In Experiment 2, we confirm that vowel duration is cue to the contrast and that perceptual categorization of vowel duration itself is also mediated by contextual phrasing information. Results thus suggest that prosodic phrasing, signaled by intonation only, mediates perception of the contrast with temporal context controlled. We discuss these findings in terms of their implications for the role of phrasing in segmental perception and in processing.

Keywords: speech perception, prosody, prosodic phrasing, Korean, Accentual Phrase

1. Background

The perception of cues to phonemic contrasts in language is well known to be highly context-dependent. One type of contextual influence which has received recent attention in the literature is prosodic structural context, including phrase-level prosodic organization (Kim and Cho, 2013; Kim et al., 2018a; McQueen and Dilley, 2020; Mitterer et al., 2016, 2019; Steffman, 2019a, 2020, 2021b). The relevance of prosodic information becomes apparent in light of a body of literature showing that phrasal organization fine-tunes speakers' articulation of segmental material, systematically modulating acoustic cues to a given contrast, which we describe in more detail below (e.g., Cho, 2005, 2015, 2016; Keating et al., 2004; Keating, 2006). These patterns motivate the question of if and how prosodic structural information is brought to bear on listeners' interpretation of segmental cues, that is, acoustic cues which are used by listeners in their perception of segmental categories.

Here and throughout we use the term *prosodic context* to refer to listeners' interpretation of prosodic structure in speech, for example, the presence of a prosodic boundary (reflecting phrasing). Prosodic context will thus be signaled by acoustic cues such as duration, f0, and voice quality. We use the term *temporal context* to refer to the timing and duration of speech sounds reflecting, for example, the speech rate of an utterance. Since prosodic context often, though not always, modulates the temporal realization of segments at a prosodic juncture (e.g., Byrd and Saltzman, 2003; Cho, 2015), prosodic context can be considered to be correlated with temporal context. For example, a relatively lengthened segment may be caused by the presence of an upcoming phrasal boundary, marked via phrase-final lengthening. However, a lengthened segment, which cues prosodic context, can also be described as a local slow-down in speech rate; in this sense it is also a piece of temporal context.

This relationship between temporal and prosodic context has recently raised a thorny issue in the literature: whether listeners' shifting perception of cues to segmental contrasts based on prosodic context comes about as a necessary consequence of computing a high-order prosodic structure (using any relevant temporal/spectral cues that signal it).

This issue is particularly pertinent for the perception of temporal segmental cues, for example, the duration of voice onset time (VOT) as a cue to voicing contrasts. Temporal segmental cues are perceived as a function of low-level temporal context, a process often referred to as "speech rate normalization". Temporal and prosodic context often make the same predictions about how listeners should shift their perception of a contrast(Mitterer et al., 2016). Demonstrating an independent effect of prosodic context would thus better our understanding of the types of information (e.g., low-level temporal information versus higher-order prosodic-structural information) which listeners integrate in segmental perception, and bear on related processing questions (Kim et al., 2018b; Mitterer et al., 2019). The present study addresses this outstanding question, adopting the test case of a small phrasal domain in Seoul Korean.

1.1. Prosodic context or temporal context?

VOT in stops varies in a systematic way based on prosodic phrasing, generally following a pattern of "initial strengthening": at the initial position of higher-level prosodic domains, as compared to the initial position of lower-level prosodic domains, VOT is longer. Relatedly, VOT at the beginning of a phrasal domain is longer than medial to that domain. For example, Intonational Phrase (henceforth IP) initial VOT is longer than IP-medial VOT. This is the case both for aspirated stops with long lag VOT (Keating, 2006; Keating et al., 2004), and in some cases for unaspirated stops with short lag VOT (e.g., word-initial /b,d,g/ in American English, as shown in Kim et al., 2018a). Given this, we might expect that listeners would integrate information about prosodic context (here specifically, prosodic boundaries) in their perception of VOT. Indeed, Kim and Cho (2013) found that when a stop from a /p/-/b/ continuum was preceded by an IP boundary, overall longer VOT was required for an aspirated /p/ percept: listeners in a sense "expected" longer post-boundary VOT, and brought this to bear on their perception of VOT as a cue to the contrast. In this sense, listeners' parse of the prosodic structure of an utterance (computed based on various cues such as duration and f0), is hypothesized to be integrated as a piece of contextual information in segmental and lexical processing (see e.g., Cho et al., 2007;

Mitterer et al., 2019 for discussion along these lines). The effect in Kim and Cho (2013) could in this sense be described as a prosodic context effect. Results such as these have motivated the proposal that language processing involves parallel computation of prosodic and segmental material and interaction between these two domains (Mitterer et al., 2019; McQueen and Dilley, 2020; Steffman, 2020). As stated by Kim et al. (2018b, p. 26) "[. . .] prosodic structure is parsed in parallel to the segmental level and is used later for prosodic modulation in lexical access".[1]

However, this same effect may also be explained only by temporal context: the acoustic duration of segments surrounding the to-be-categorized stop. Because phrase-final lengthening serves as a cue to the presence of an IP boundary, an IP-initial stop preceded by an IP-boundary is thus also preceded by a segment that has undergone phrase-final lengthening (and is longer in duration as compared to an IP-medial equivalent). This should be considered in light of the fact that listeners' perception of temporal segmental cues in speech (e.g., VOT) is highly sensitive to temporal context, in particular the well-studied mechanism of "durational contrast" (e.g., Diehl and Walsh, 1989; Wade and Holt, 2005). Durational contrast refers to a perceptual effect whereby the "perceived length of a given acoustic segment is affected contrastively by the duration of adjacent segments" (Diehl and Walsh, 1989, p 2154), and is also often referred to as "speech rate normalization". Consider the implication of this effect for the aforementioned case of stop VOT: longer pre-VOT segment durations (as a cue to a preceding IP boundary) would contrastively impact perception of VOT such that it sounds relatively short, more like /b/. Like an account which implicates prosodic structure, this durational contrast (temporal context) account predicts that listeners should require overall longer VOT to perceive aspirated /p/ when a longer

[1]This proposal of "prosodic analysis", as originally formulated in Cho et al. (2007), postulates that listeners use both prosody and segmental information in word recognition along these lines, with segmental material activating lexical candidates and prosodic-structural information being integrated in lexical competition, e.g., in the alignment of word boundaries and phrasal boundaries (see also, e.g., Christophe et al., 2004).

vowel precedes the stop (i.e. when the stop is IP-initial).

In summary, we have two possible accounts of the same data. In one, a prosodic structural representation, which is computed on the basis of information that signals prosodic structure (e.g., phrase-final lengthening), informs the process of mapping cues to segmental structure. In the other, perception shifts as a function of temporal context (without reference to prosody), in line with well-documented contrast effects. Durational contrast and other speech rate effects are typically thought to result from low-level, or general auditory processing, being immune to changes in cognitive load and selective attention (Bosker et al., 2017, 2020), occurring for non-speech stimuli (Wade and Holt, 2005), and in non-human species (Lotto et al., 1997). These two accounts thus implicate very different processing mechanisms, and more specifically, the extent to which prosodic structural information guides the interpretation of temporal cues in speech.

A possible solution is to manipulate other cues to prosodic context, without creating temporal variation. One promising cue in this regard is pitch, as related to the intonation of a language, that is, f0 tunes signaling prosodic structure, including phrasing (Jun, 2005, 2014; Ladd, 2008). Following this reasoning, Steffman (2019a) tested if an intonational pattern, which occurs (unambiguously) at IP boundaries, would shape listeners' perception of post-boundary cues in American English. Steffman (2019a) employed a similar design to Kim and Cho (2013), testing perception of a VOT continuum. However, in lieu of serving as a direct cue to a prosodic boundary, Steffman found that an unambiguous IP boundary tone sequence, when presented without phrase-final lengthening, had the unintended consequence of modulating listeners' perception of speech rate. In other words, a boundary tone overlaid on a short vowel was interpreted as an increase in speech rate, due to the typical co-occurrence of IP boundary-marking tones and (local) rate slowdowns (i.e. phrase final lengthening). Thus complicating the picture, tonal cues to prosodic structure can alter perceived speech rate, and interpretation of durational cues (in lieu of directly cuing a prosodic boundary). Larger phrasal domains like the IP are marked both by tonal events and temporal changes at their right edge, so that manipulating pitch alone is likely

to (and perhaps will necessarily) modulate perception of speech rate.

2. The present study

As we have laid out, the prevalence of lengthening as a cue to a prosodic boundary (1) offers it as a potential explanation for perceptual shifts in some cases as discussed by Mitterer et al. (2016), and (2) further complicates the use of pitch-based cues as shown in Steffman (2019a). Both of these issues are related to the fact that IP boundaries are typically marked with robust lengthening at the right edge. However, for smaller prosodic phrases, this is not necessarily the case.

The present study was thus designed to test for an independent influence of phrasing in listeners' perception of durational cues by examining the Accentual Phrase (henceforth AP), a small prosodic unit slightly larger than a word (see e.g., Jun, 2005, 2014). The AP is defined primarily by certain tonal patterns. Unlike the Intonational Phrase in Korean (and many other languages), there is not robust phrase-final lengthening at the AP's right edge (Jun, 1995, 1998; Koo, 1986). The AP is further a domain for phonological processes (Jun, 1996, 1998; Kim et al., 2018b), and patterns of segmental modulation which are generally framed as domain-initial strengthening (as, for example, evident in an increase VOT of an aspirated stop in AP-initial position, heightening its phonetic clarity at a prosodic juncture). These patterns are described in Section 2.2.

As such, we might predict that a perceived difference in phrasing will lead to shift in the perception of said AP-initial cues. Crucially, since an AP in Korean is not marked by robust lengthening, we can cue AP phrasing with no temporal variation, and naturalistically signal an AP-boundary with pitch alone.

To summarize, studies up to this point have not succeeded in disentangling temporal and prosodic context as influences in listeners' perception of temporal cues in speech. While in some cases both influences predict the same perceptual adjustments, they implicate very different mechanisms. As such, an answer to the question "does prosodic context independently influence perception of temporal cues?" will allow us to refine our understanding of how prosody enters into speech comprehension and inform theoretical accounts of prosody

in segmental and lexical processing (Cho et al., 2007; McQueen and Dilley, 2020; Mitterer et al., 2019), which are discussed in Section 2.1.

To answer these questions, we set out in this study to test how two temporal segmental cues to a Korean stop contrast, VOT and vowel duration, are perceived as a function of prosodic phrasing at the level of the AP. Below some key facts about Korean prosody and the relevant segmental contrasts are outlined.

2.1. The AP and prosodic phrasing in speech processing

Most relevant for our purposes are two properties of the AP. First, the AP in Seoul Korean is delimited by f0 on its left and right edges. In the prevalent Seoul Korean intonational phonology model (Jun, 1993) the default tonal pattern is THLH where the AP-initial tone "T" can be L or H based on the laryngeal feature of the AP-initial segment. The AP-initial tone is H if the AP-initial segment is aspirated or fortis, but L otherwise.[2] The AP-second tone, H, is typically realized on the 2nd syllable of an AP, but can be realized on the third syllable when an AP is longer than 4 syllables (Jun, 2000; Jun and Cha, 2015). When an AP contains fewer than four syllables, the second and the third tone may be subject to undershoot. The AP-final tone tends to be consistently marked by the AP-final H tone (or, a final rise) although it may rarely be realized as an L tone due to tonal interactions in some cases.

Secondly, as noted above, AP right edges are not marked by temporal modulations. As stated by Jun (1998, p 193):

> Segment duration [...] is not greater at the end of an AP (Koo, 1986; Jun, 1995). This differs from the duration pattern of an IP: an IP-final syllable is lengthened

[2]However, this segment-tone mapping, being postlexical, is sometimes violated in natural conversation without affecting the meaning of the sentence, as described in Jun (2000). Based on speech data from reading, acting, and interviews, Kim (2004) found that AP-initial segment-tone mapping was violated about 7% of the time when the AP begins with H-tone triggering segments and about 5% of the time when the AP begins with L-tone triggering segments.

> substantially at the end of the phrase (Jun, 1993) [...].[An AP boundary] is never followed by a pause unless it is the last phrase of an IP.

These properties suggest that raised f0 should provide a strong cue for an upcoming AP boundary, without modulating listeners' perception of localized speech rate at a prosodic juncture. Indeed, tonally cued Korean AP boundaries have been shown to be important in speech processing. One clear impact is evidenced in word segmentation where the presence of AP boundary cues, including an AP-final H tone, facilitates word segmentation in word spotting experiments (Kim, 2004; Kim and Cho, 2009). In addition to segmentation, AP-phrasing has been shown to mediate phonological inferencing, based on the AP's function as a domain for the application of phonological processes (Kim et al., 2018b).

2.2. Korean laryngeal contrasts and AP-initial strengthening

Here we outline the relevant effect of AP phrasing on two temporal cues to Korean laryngeal contrasts. The AP serves as a domain for various patterns of "strengthening", in which segmental properties are enhanced domain-initially (Jun, 1993, 1996; Cho and Keating, 2001; Keating et al., 2004). The contrast we chose to test is between aspirated and fortis stops.[3] Two cues which distinguish these stops are VOT and vowel duration, where longer VOT (Cho and Jun, 2000; Cho and Keating, 2001) and shorter vowel duration (Cho, 1996; Chung et al., 1999; Choi, 2011a; Kim, 2002) are observed for aspirated, as compared to fortis stops. Choi (2011a) further shows that vowel duration differentiates the

[3]The Seoul dialect of Korean has a three-way laryngeal contrast in obstruents, with lenis, fortis and aspirated categories. When produced AP-initially, pitch has recently become a major cue for the contrast between lenis and aspirated stops in younger speakers (Kang, 2014), with the VOT cue merging between these two stop categories. Fortis and aspirated stops are not distinguished by pitch.

stop categories in this way in both utterance-initial and utterance-medial position.[4] Both of these cues are additionally used by listeners in perception of the contrast (Choi, 2011b; Kim et al., 2012).

Laryngeal contrasts in Korean have been argued to undergo paradigmatic strengthening in phrase-initial position. For example, Cho and Keating (2001) and Cho and Jun (2000) found that Korean aspirated /t^h/ undergoes systematic lengthening of VOT corresponding to its position in the prosodic hierarchy such that word-initial VOT is shorter than AP-initial VOT which is shorter than IP-initial VOT: being initial to a higher-level prosodic domain "strengthens" a stop's specification as aspirated. In comparison, VOT in fortis stops does *not* systematically lengthen at higher-level domains, and remains relatively short. In fact, Cho and Jun (2000) found that VOT for fortis stops is slightly shorter at higher-level domains, effectively enhancing the contrast between aspirated and fortis stops in terms of VOT.

Strengthening of this sort is also evident in the duration of the following vowel in a domain-initial CV syllable: Cho and Keating (2001) found that post fortis stop vowels undergo lengthening AP-initially (as compared to AP-medially), where 97% confidence intervals for measurements of AP-initial versus word-initial vowel durations in fortis stops do not overlap. This vowel lengthening enhances the relationship between shorter preceding VOT and a longer following vowel as a cue to fortis stops. On the other hand, the vowel duration after the AP-initial aspirated stop was not as lengthened as it was for the fortis stop.

[4]Cho and Keating (2001) notably document a different relationship between vowel duration and stop category. In their data, aspirated stops show longer post-stop vowel duration than fortis stops, which differs from the other literature measuring vowel duration for these stops. This discrepancy is likely explained by the syllable structure of the target items used in Cho and Keating (2001), where the word used to elicit tense stops was /t*ak.pu.ri/ (a nickname), with a closed first syllable. The word used to elicit aspirated stops was /t^ha.dʒa.ni/ "Tarzan", with an open first syllable (this is a loanword and the first vowel is also likely to be lengthened following a loanword pronunciation rule). Because vowels in closed syllables are generally shorter than those in open syllables in Korean (Choi and Jun, 1998), the syllable structure of the items used is a likely explanation for the difference between Cho and Keating (2001) and the other studies cited here.

While mean vowel duration after aspirated stops was longer AP-initially than AP-medially, 97% confidence intervals reported in Cho and Keating (2001) overlapped substantially.

Both vowel duration and VOT accordingly become more different from one another for these two stop categories in AP-initial, as compared to AP-medial (word-initial), position. In this sense, the temporal relationship between VOT and vowel duration as a cue to each stop category is enhanced domain-initially. It is worth bearing in mind that these patterns of strengthening are asymmetrical: VOT in aspirated stops lengthens robustly domain-initially, but does not change substantially for fortis stops, whereas the opposite is true for vowel duration (vowel duration after fortis stops lengthens robustly domain-initially but does not change substantially after aspirated stops). Given the patterns outlined above, our goal will be to test how prosodic context mediates listeners' perception of VOT and vowel duration as cues to the stop contrast. We carry out two experiments to address this question. In Experiment 1, we test listeners' perception of the fortis-aspirated stop contrast based on VOT, while manipulating f0 as a cue to AP phrasing. Building on the results of Experiment 1, in Experiment 2 we test how both vowel duration and VOT, when varying orthogonally, contribute to perception of the contrast, and how perception of vowel duration additionally shifts as a function of phrasing.

3. Experiment 1

In Experiment 1 we tested how listeners' perception of a VOT continuum shifted depending on the location of stops within an AP, in line with how VOT varies as a function of AP-initial strengthening of the contrast between fortis and aspirated stops. The contextual manipulation, described below, crucially held temporal context constant, eliminating it as possible explanation for an observed shift in categorization of the continuum.

3.1. Materials

Our goal in designing the materials for Experiment 1 was to create a VOT continuum ranging from a fortis velar stop /k*/, with relatively short VOT, to an aspirated velar stop /k^{h}/ with longer VOT. This target sound was cued as either AP-medial or AP-initial in a

carrier phrase. The carrier phrase was chosen to be natural if phrased in either way, and translated to "that's what we inserted/grew ", where the meaning of the verb is changed by whether it begins with a fortis stop ("to insert") or aspirated stop ("to grow").[5] These two conditions are shown in (1) and (2) below, transcribed in IPA, with parentheses indicating AP phrasing (curly brackets indicate IP phrasing). The question mark indicates the target sound which will be categorized as /k^{h}/ or /k*/.

(1) AP initial: {(kɯkʌ)$_{AP}$}$_{IP}$ {(ulika)$_{AP}$ (?iwʌs*ʌ)$_{AP}$}$_{IP}$
gloss: *that* *we* *inserted/grew*

(2) AP medial: {(kɯkʌ)$_{AP}$}$_{IP}$ {(ulika ?iwʌs*ʌ)$_{AP}$}$_{IP}$

In (1) the target segment, marked by '?', is preceded by an AP boundary and is therefore AP-initial, while in (2) it is phrased with the preceding word and is therefore AP-medial. For the reasons outlined above, our goal was to create this perceived difference in phrasing by manipulating only f0 in the carrier phrase. The contextual manipulation is shown in Figure 1.[6]

A female native speaker of Seoul Korean in her twenties read the sentence with /k^{h}/ and /k*/ multiple times, and with each of the phrasing patterns above. The speaker was asked to read the sentence as naturally as possible.[7] The tokens with and without an AP boundary between the second and the third word ("we" and "inserted"/"grew") were selected to be used as references for manipulation, corresponding to (1) and (2) above. The recording was

[5]In *Hangul* orthography: 그거 우리가 끼웠어/키웠어.

[6]The f0 values in medial pre-target /ka/ are relatively high in our stimuli. As noted in Section 2.1, this can occur when an AP contains more than four syllables as it did in the medial condition (Jun, 2000; Jun and Cha, 2015).

[7]The speaker was not trained in K-ToBI and was not aware of the purpose of the study. Before the recording, a K-ToBI trained experimenter produced the sentence with two different phrasing patterns (corresponding to (1) and (2) above) and asked the speaker to repeat after the model production. The speaker was able to reproduce the model phrasing without any difficulty. She then read the sentence with two different phrasing patterns multiple times during the recording session.

made in a sound-attenuated booth, using a Tascam HP-D2 digital recorder and a SHURE KSM44 microphone at a sampling rate of 44 kHz.

The manipulation of stimuli was carried out by re-synthesizing f0 and duration of naturally produced utterances, using the PSOLA method (Moulines and Charpentier, 1990), as implemented in Praat (Boersma and Weenink, 2020). The base files for this process were two utterances which corresponded to (1) and (2) above, produced with an aspirated stop /k^{h}/. We started with an utterance that was produced with an aspirated stop in AP-medial position, as in (2). We examined the duration of the post-stop vowel /i/ and manipulated it to be the average of this production (25 ms) and that produced following a fortis stop /k*/ (75 ms), resulting in a vowel duration of approximately 50 ms. In this way, vowel duration should be relatively ambiguous, encouraging listeners to use VOT as a cue to the contrast. We then manipulated VOT to vary in 7 ms increments, ranging from 30 ms, corresponding to a /k*/ endpoint, to 79 ms corresponding to a /k^{h}/ endpoint, for a total of 8 steps, which were based on approximate durations for each stop category as produced by the model speaker. These endpoints were judged to sound like clear productions of each stop category by three native Korean speakers (authors SK, TC and SJ).

Next, we manipulated AP phrasing by re-synthesizing f0 on the immediately pre-target syllable in the carrier phrase, i.e., /ka/. We resynthesized f0 in both conditions, so that each underwent the same amount of manipulation. To create the AP-medial condition, we overlaid highly comparable f0 from another AP-medial production on the base file. To create the AP-initial condition, the f0 from an AP-initial production (with an AP boundary preceding the target) was overlaid on the pre-target syllable. The f0 contour on the pre-

[8]The AP-initial syllable beginning with the target stops (aspirated or fortis) did not have high f0, but as mentioned earlier, violations of the AP-initial tone-segment mapping rule occur occasionally, and do not affect meaning. As stated by Jun (2000, p 4): "[...] none of the surface tonal variations which deviate from the underlying tonal sequence seem to have a different meaning. What is meaningful in Korean intonational phonology is the phrasing, marked by the boundary tone of an AP or an IP." The low f0 on the target syllable with an aspirated/fortis onset is still perceived as AP-initial because the f0 on the pre-target syllable is the highest in the phrase, thus clearly cuing an AP-final H (cf. Kim and Cho, 2009).

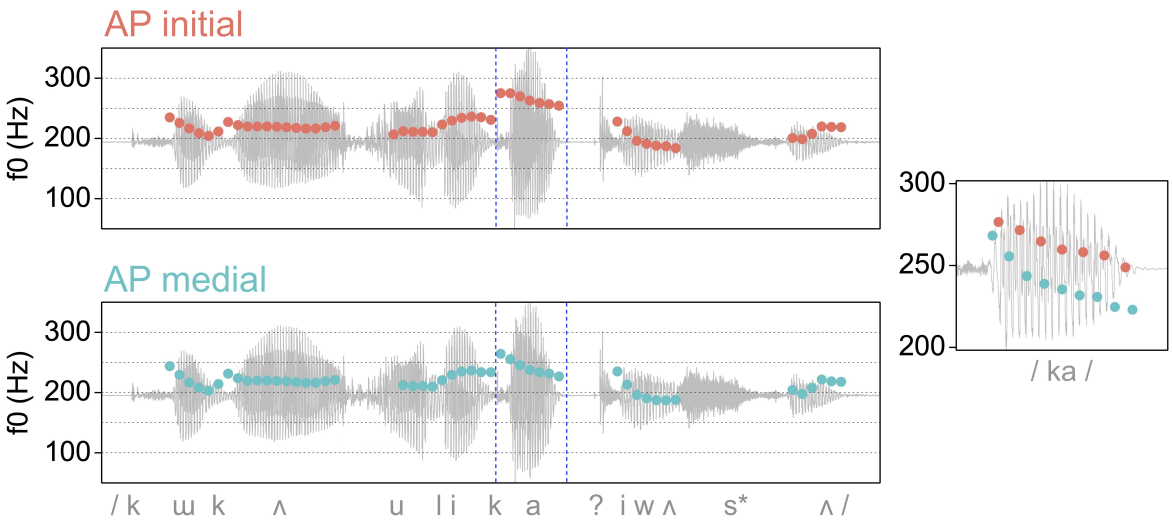

Figure 1: Example waveforms, with overlaid pitch tracks for stimuli, showing the phrasing manipulation used in both Experiments. The critical pre-target syllable /ka/ is highlighted by the dashed blue lines, and both pitch tracks for this syllable are plotted together at right for comparison. A segmental transcription is given below the waveforms, with ? representing the target sound. The 30 ms VOT step from the continuum is shown here.

target syllable was clearly higher as compared to AP-medial f0, as shown in Figure 1. Notably, f0 on the target itself was left un-manipulated with the AP medial values, as were all other syllables in the carrier phrase.[8] This was judged by authors SK, TC and SJ to sound like an appropriate pitch contour for the target syllable in each AP phrasing, which allowed for it to remain identical across conditions. This kept our contextual manipulation very minimal, while still conveying a clear difference in phrasing. These manipulations resulted in a total of 16 unique stimuli (8 continuum steps × 2 contexts).

3.2. Predictions

Given the structure of the stimuli, and the ways in which the contrast between fortis and aspirated stops is strengthened AP-initially, we can entertain two predictions.

First, we can predict that, overall, there will be decreased aspirated /k^{h}/ responses in the AP-initial condition if listeners are expecting AP-initial lengthening of VOT. As

described above, both VOT and vowel duration in aspirated and fortis stops are modulated by phrasing at the level of the AP. Aspirated stop VOT shows systematic lengthening AP-initially, while fortis stop VOT does not change systematically, though it may shorten slightly (Cho and Jun, 2000; Cho and Keating, 2001). If listeners are sensitive to the VOT lengthening pattern, they should thus overall require longer VOT to perceive an aspirated stop in the AP-initial condition, as compared to the AP-medial condition. Note that given the lack of a systematic change for fortis stops, we don't have a clear basis for making a competing prediction based on that pattern.

Second, we can make a prediction based on the vowel in the syllable containing the target stop. Recall that vowel duration is lengthened AP-initially, more robustly for fortis stops, and less so for aspirated stops (Cho and Keating, 2001). Vowel duration is also longer in fortis stops as a cue to the contrast for listeners (Choi, 2011b). Recall that in the stimuli, the duration of the post-stop vowel was manipulated to relatively ambiguous. Because vowel duration needs to be longer for the percept of a fortis stop, especially in the AP-initial position in which the following vowel is usually longer than in the AP-medial position, the post-stop vowel should sound shorter than would be expected in an AP-initial, post-fortis-stop context. In other words, if listeners expect AP-initial lengthening of the (ambiguous) post-stop vowel, and bring this to bear on their perception of the contrast, the vowel should sound relatively *short* in the AP-initial condition. This relatively shorter perceived vowel duration would cue an aspirated stop, predicting increased aspirated responses in the AP-initial condition. It is important to note that, given that we are manipulating only VOT in Experiment 1, this latter result should be evident in cases where VOT values are fairly ambiguous such that vowel duration becomes a pertinent cue.

We thus have two possible outcomes which would reflect listeners' sensitivity to prosodic context in their perception of the stop contrast. The directionality of the effect might further tell us which cue is most relevant in this regard (VOT or vowel duration). Importantly, since the duration of the surrounding speech material before and after the test word remains the same across conditions (AP-initial and AP-medial), any shift in perception of the fortis-

aspirated stop contrast would not be attributable to changes in temporal context.

3.3. Participants and procedure

We recruited 30 native speakers of the Seoul Dialect of Korean, born and raised in the Seoul and Gyeonggi area, and residing in the area at the time of the experiment (25 self-identified females, 5 males; age range 20-30, mean age 22.7). Participants were asked to take part in the online experiment remotely, on their personal computer in a quiet room, wearing headphones. Three additional participants took part in the experiment, but were excluded from analysis because they did not wear headphones as instructed, or were not from the Seoul and Gyeonggi area. All participants were undergraduate students in their 20s, and were located in Seoul at the time of data collection. They were paid for participation in the study. Experimental stimuli were presented online using the platform Appsobabble (Tehrani, 2020).

Participants were instructed that their task was to listen to speech and to select which word they heard spoken. As part of the experiment instructions, they were shown the sentence they would be listening to orthographically, and an example of the trial display, in which the choice of the critical word was displayed in *Hangul* orthography on the computer monitor, each choice centered on each left/right half of the computer screen.

The procedure was a simple two alternative forced-choice (2AFC) task in which participants listened to an auditory stimulus and indicated their response by using the computer keyboard. They indicated their response using the 'f' and 'j' computer keys, which corresponded to the choice on the left side of the screen and the right side of the screen, respectively. After a categorization response was made via key press, the next trial began automatically, with a short (200 ms) interval between the key press response and the onset of the next trial. In each trial the stimulus was only played once, and could not be replayed. Each of the 16 unique stimulus were presented a total of 14 times each in randomized order to participants for a total of 224 trials. There were no practice trials, and all responses were analyzed. The experiment took approximately 15-20 minutes to complete.

3.4. Statistical modeling

We assessed our results statistically using a log-link Bayesian mixed effects model implemented using brms (Bürkner et al., 2017), as implemented with R (v 4.1.2, R Core Team, 2021), in the RStudio environment (RStudio Team, 2020).

The model predicted listeners' categorization response (/k*i/ versus /k^hi/, with /k^hi/ mapped to 0 and /k*i/ mapped to 1) as a function of VOT (centered and scaled) and phrasing context (contrast coded with AP-initial mapped to 0.5, and AP-medial mapped to -0.5), and the interaction of these two fixed effects. The random effect structure of the model contained intercepts for participant and by-participant random slopes consisting of the same structure as the fixed effects (in model code: response ~ VOT*context+(1+VOT*context | participant)). We used weakly informative priors, implemented as normal(0,1.5) in log-odds space, both for the intercept and for fixed effects. The data frames with experimental data, and the code for analysis is available in an open access repository hosted on the Open Science Foundation at `https://osf.io/72asn/`.[9]

In assessing the impact of a given fixed effect on categorization, our interest is in observing if the estimated credible intervals exclude zero, suggesting that the estimate is reliably non-zero and has a consistent directionality: a reliable impact on categorization. We also report the proportion of the posterior which shows a given sign, as calculated with the `p_direction` function in the R package bayestestR (Makowski et al., 2019b). This metric has an interpretation which is more intuitively compared to a frequentist p-value, and corresponds fairly closely with it (Makowski et al., 2019a). For example, if 99% of the posterior distribution shows a given sign (positive or negative), we can infer a 99% probability that the effect does indeed show this directionality. We would indicate this as pd = 0.99.

[9]Due to the imbalanced sample of participant gender, we also ran a model which included gender as a fixed effect, and interaction term with all other fixed effects. The same was done for Experiment 2, which also contained an imbalanced gender sample. These models found no evidence for a main effect of gender or an interaction of gender with any of the other fixed effects. These gender-including models are not discussed here, though the full models and their summaries can be found on the open access repository.

3.5. Results and discussion

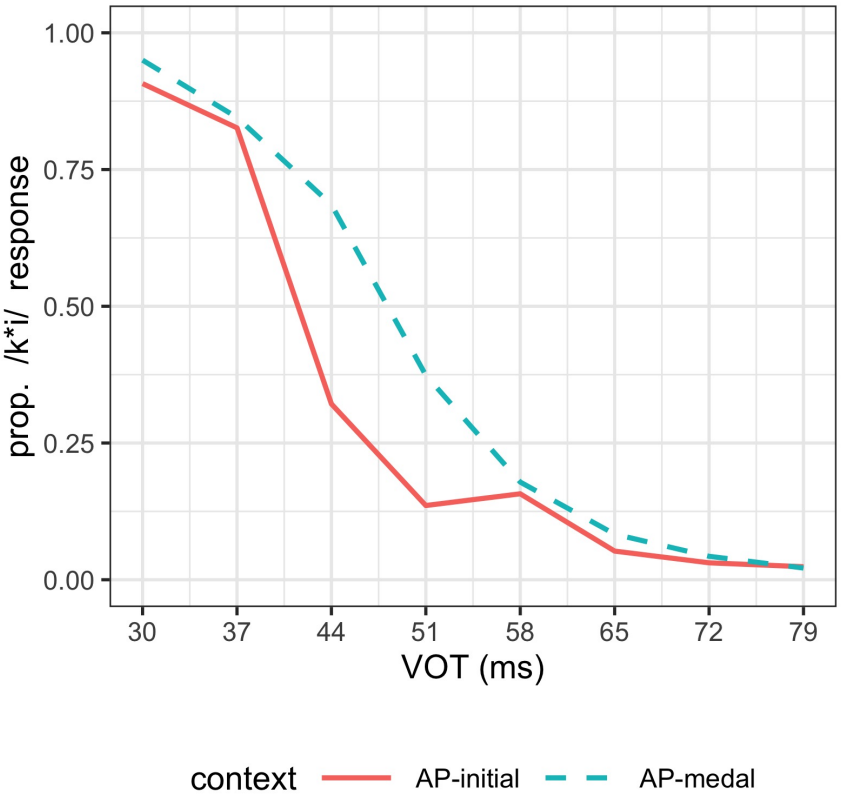

Figure 2: Categorization responses in Experiment 1 along the continuum, shown on the x axis, with the y axis plotting the proportion of fortis /k*i/ responses. Categorization is split by phrasing manipulation, indicated by line type and color.

The model output for fixed effects is given in Table 1, which provides the marginal posterior distribution for fixed effects in the form of the median of the estimated distribution as well as 97.5% credible intervals for an estimate. As expected, changing VOT along the continuum evidenced a credible impact on responses whereby as VOT increased, fortis responses decreased (β=-2.99, 97.5%CI = [-3.53,-2.45]; pd = 1), in line with longer VOT cuing aspiration. The intercept in the model was also credibly non-zero, showing a bias towards aspirated responses overall along the continuum. This may be due to the fact that the basis for the creation of the stimulus was an aspirated stop, which likely contained

additional cues to the contrast, such as changes in voice quality (Cho et al., 2002). We note that nevertheless continuum endpoint responses are fairly well-anchored.

Phrasing context, the main predictor of interest, also evidenced a credible impact on categorization (β=-1.12, 97.5%CI = [-1.53,-0.74]; pd = 1), showing *decreased* fortis responses in the AP-initial position, relative to AP-medial position. Recall in our predictions that listeners' expectation of AP initial lengthening of VOT would predict that longer VOT should be required for an aspirated response in the initial condition. In other words, listeners should more readily categorize the target as being a *fortis* stop, predicting increased fortis responses in the AP-initial condition, the opposite of what we observe. Instead, this outcome is consistent with the prediction that listeners expect AP-initial lengthening of the post-stop vowel, such that in AP-initial position this vowel sounds relatively short. A shorter perceived vowel in the AP-initial condition, if used as a cue to the stop contrast, would signal an aspirated stop, thus decreasing fortis responses therein. To the extent that this interpretation is correct, we have found evidence that, independent of temporal context, listeners are integrating prosodic phrasing information with their perception of temporal cues.

This interpretation is in line with the fact that we only see the f0-based prosodic context effect when categorization is not well anchored (i.e. VOT values are ambiguous), such that vowel duration becomes more relevant as a cue to the contrast. However, because we did not actually manipulate vowel duration in this experiment, we cannot be sure that listeners are indeed using it as a cue; we also cannot be certain of the way in which vowel duration trades as a cue to stop category with VOT in our stimuli. Accordingly, we implemented

Table 1: Model estimates for fixed effects in Experiment 1

	Estimate	Est. Error	L-97.5% CI	U-97.5%CI	credible?
intercept	-1.64	0.21	-2.11	-1.15	✓
context	-1.12	0.17	-1.53	-0.74	✓
VOT	-2.99	0.24	-3.53	-2.45	✓
VOT:context	-0.07	0.19	-0.51	0.34	

Experiment 2 to address these questions and replicate our finding in Experiment 1.

4. Experiment 2

The goal of Experiment 2 was to test the same contextual influence which was seen in Experiment 1, while also varying vowel duration following the target stop to examine (1) if/how vowel duration cues the fortis-aspirated stop contrast in our stimuli and (2) how categorization of vowel duration-varying stimuli is impacted by AP-phrasing context.

4.1. Materials

The materials from Experiment 2 were created by adding a manipulation of vowel duration to the materials used in Experiment 1. This was accomplished by manipulating the duration of the vowel [i] following the target stop. The duration of this vowel in Experiment 1 was approximately 50 ms, and in Experiment 2 it was manipulated to range from 20 ms to 80 ms in 20 ms steps, for a total of four. This range was judged to sound natural given the overall speech rate of the utterance. We additionally added one more VOT continuum step, shortening VOT to be 23ms for the fortis endpoint in the continuum with the goal of eliminating the slight bias towards aspirated responses seen overall in the continuum in Experiment 1. These manipulations are shown schematically in Figure 3, where the fortis endpoint of the continuum is shown in the top left of the figure and the aspirated endpoint is shown in the bottom right. These manipulations resulted in a total of 72 unique stimuli (9 VOT steps $\times$ 4 vowel duration steps $\times$ 2 phrasing contexts).

4.2. Participants and procedure

30 participants were recruited for Experiment 2, none of whom had participated in Experiment 1 (25 self-identified females, 5 males; age range 22-29, mean age 24.6). All were undergraduate students, born and raised in the Seoul and Gyeonggi area, and located in Seoul at the time of data collection. Nine additional participants took part in the experiment, and seven were excluded from analysis because they did not wear headphones as instructed, or were not from the Seoul and Gyeonggi area. Two of these nine additional

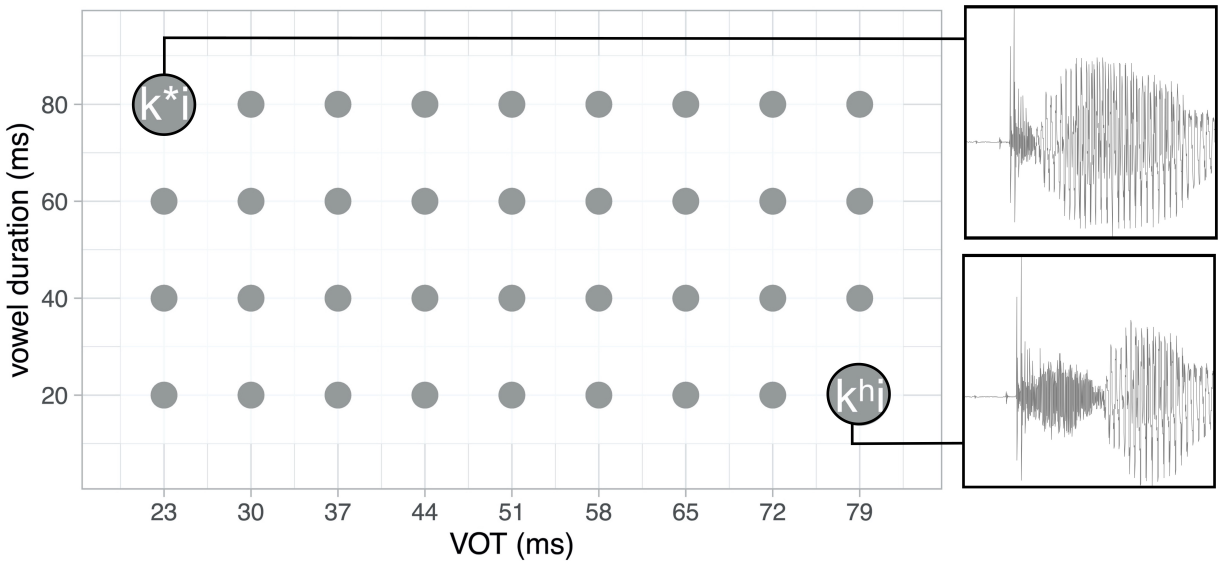

Figure 3: A representation of the two dimensional continuum used in Experiment 2, with VOT on the x axis and vowel duration on the y axis. Continuum endpoints are circled and shown as waveforms at right.

participants were excluded because categorization responses at continuum endpoints were under 50% of the expected categorization response, suggesting the continuum was not perceived as the intended stop categories (there were no participants like this in Experiment 1). The procedure and instructions were identical to Experiment 1. Each of the 72 unique stimuli was presented 3 times with all stimuli randomized, for a total of 216 trials in the experiment. The experiment took approximately 15-20 minutes to complete.

4.3. Statistical modeling

We assessed our results with the same statistical modeling as in Experiment 1, however, this time we also included (scaled) vowel duration in the model as fixed effect, interacting with other fixed effects, and as a by-participant random slope, with the same general random effect structure as in Experiment 1 (In model code: response ~ VOT*vowel duration*context+(1+VOT*vowel duration*context | participant)).

4.4. Results and discussion

The model summary is shown in Table 2. Before turning to the effect of context, we will consider how VOT and vowel duration were used by listeners. Both cues were observed

Table 2: Model estimates for fixed effects in Experiment 2. "vdur" indicates vowel duration.

	Estimate	Est. Error	L-97.5% CI	U-97.5%CI	credible?
intercept	-0.06	0.14	-0.37	0.25	
VOT	-2.53	0.20	-3.00	-2.08	✓
vdur	0.67	0.07	0.52	0.83	✓
context	-0.22	0.09	-0.45	-0.01	✓
VOT:vdur	0.19	0.07	0.04	0.37	✓
VOT:context	0.00	0.11	-0.24	0.26	
vdur:context	0.01	0.08	-0.16	0.19	
VOT:vdur:context	0.02	0.11	-0.23	0.30	

to have credible influences on categorization, in line with the way they jointly signal the contrast. Increasing VOT decreased listeners' fortis responses (β= -2.53, 97.5%CI = [-3.00,-2.08]; pd = 1), while increasing vowel duration increased fortis responses (β= 0.67, 97.5%CI = [0.52,0.83]; pd = 1).[10] Both of these influences are evident in Figure 4, which shows categorization along both continuum dimensions. Though both cues have an effect, VOT is clearly a more robust cue (also suggested by the model estimates): color gradation in Figure 4 changes more strongly along the x axis, than the y axis, though it is clear that at more ambiguous VOT values (e.g., 51 ms), vowel duration has a clear influence.[11] The credible effect of vowel duration confirms our prediction that listeners clearly use it as a cue to the fortis-aspirated stop contrast in our stimuli.

[10]Of note, the intercept in the model was no longer credibly different from 0 as it was in Experiment 1, suggesting altering the VOT continuum to have one additional shorter step produced the desired result.

[11]A credible interaction between VOT and vowel duration was additionally observed. This interaction was tested in using the *emtrends* function from the package *emmeans* (Lenth et al., 2018). This test showed that VOT exerted a stronger influence at shorter vowel durations. At 20 ms vowel duration the slope estimate was -2.79, at 40 ms -2.62, at 60 ms -2.44, and at 80 ms -2.26. This is apparent visually in panel Figure 4, where gradation changes more abruptly from left to right at shorter vowel duration steps. This interaction is also visible in Figure 5.

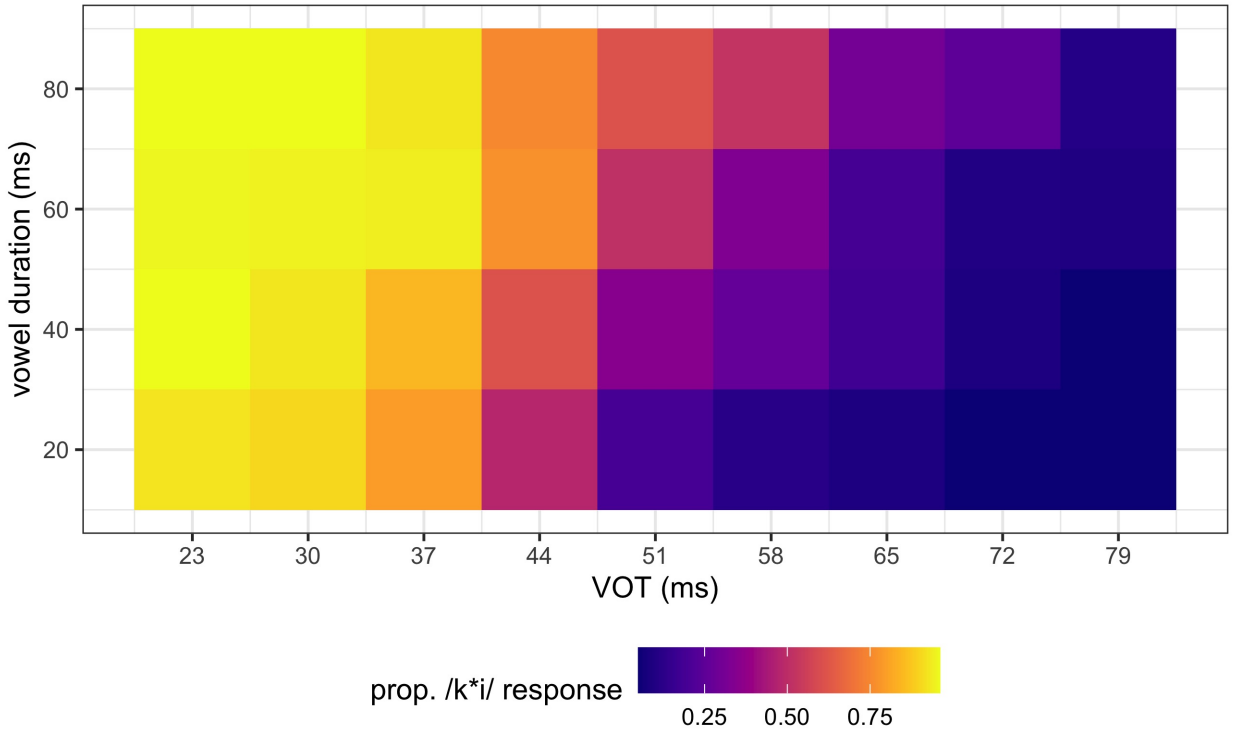

Figure 4: Categorization of the two dimensional continuum in Experiment 2, where the x axis shows VOT, the y axis shows vowel duration, and the color scale shows listeners proportion of fortis /k*i/ responses at each step (compare to Figure 3).

Turning to the effect of prosodic context, the model estimates show a replication of Experiment 1, whereby the AP-initial condition shows credibly decreased fortis responses (β= -0.22, 97.5%CI = [-0.45,-0.01]; pd = 0.99). Figure 5 shows the influence of context in two ways. First, in panel A, categorization of VOT (x axis) and vowel duration (coloration) is shown, with phrasing indicated by line type. We can see that the phrasing effect is more evident at ambiguous continuum steps, both for VOT and vowel duration, and importantly, at a fixed VOT value (e.g., 44 and 51 ms in panel A of Figure 5), context is shifting categorization of the vowel duration continuum. This can be seen in comparing same-colored lines and noting the general downward shift for solid AP-initial lines: decreased fortis /k*i/ responses in the AP-initial condition. This is particularly clear at the more ambiguous 40 and 60 ms vowel duration steps. Panel B shows a more simplified version of

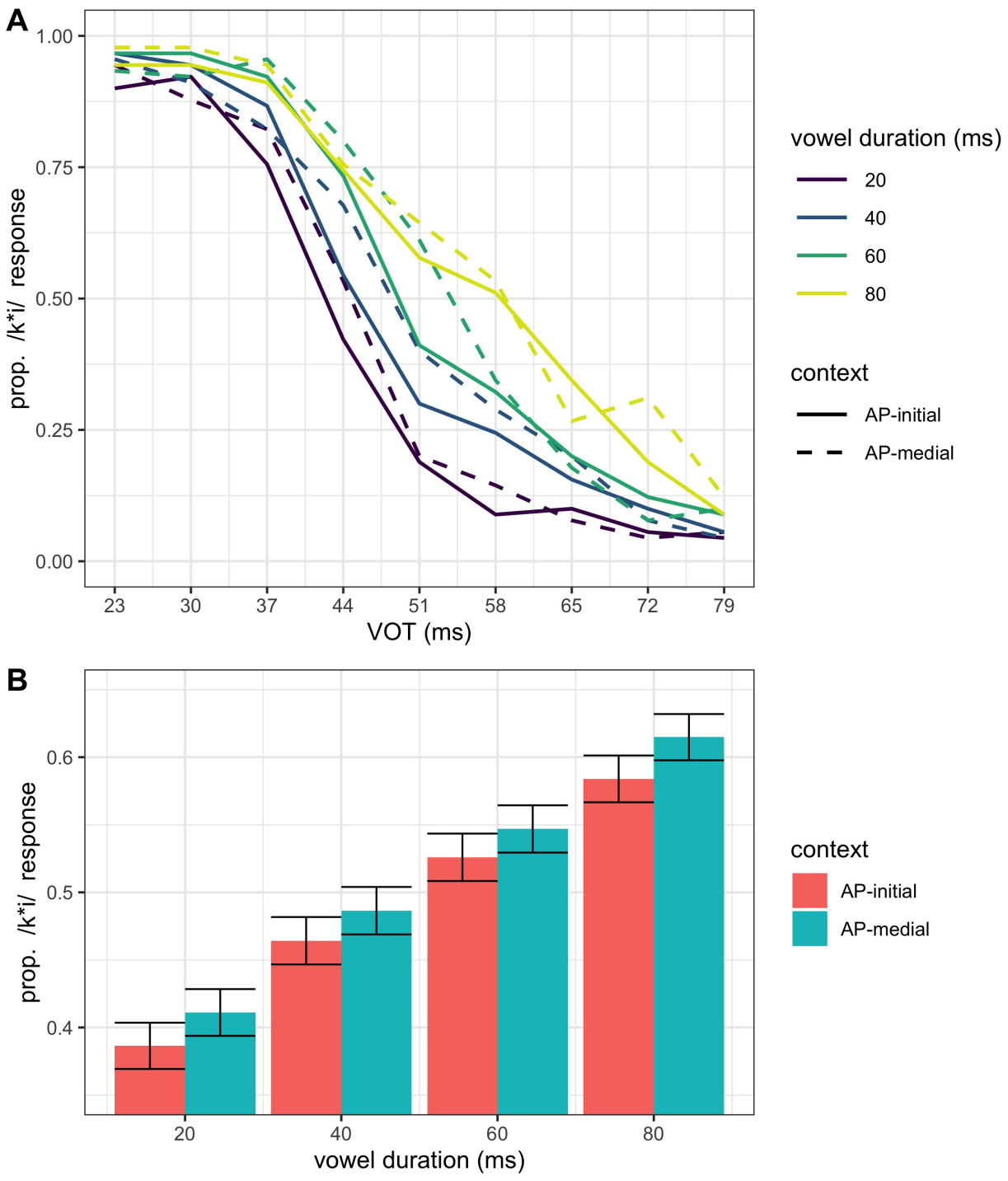

Figure 5: Panel A: Categorization of VOT (x axis) and vowel duration (coloration) split by context (line type). Panel B: Categorization of vowel duration, pooled by VOT, and split by context, with error bars showing one SE from the raw data.

the phrasing effect on vowel duration perception by collapsing across VOT for each vowel duration value. In line with panel A, we see that a given vowel duration value (pooled across all VOT values), listeners show decreased fortis /k*i/ responses in the AP-initial condition. Figure 6, included in the appendix, shows a more detailed look at categorization split by context, showing responses at each possible combination of vowel duration and VOT.

The magnitude of the contextual phrasing effect is notably smaller than what was found in Experiment 1 (β= -1.12). One possibility for the reduction of the effect size across experiments is the variation in the acoustic realization of the target present in Experiment 2 (with orthogonal vowel duration and VOT variation). Variability in a particular stimulus trait has been suggested to shift attentional resources (e.g., Green et al., 1997; Nosofsky, 1986). Green et al. (1997), who tested contextual effects along these lines, state: "Variation along a particular dimension [...] draws attention to that dimension. This increased attention may result in that dimension's being given more 'perceptual weight' than other acoustic dimensions" (p 689). With respect to prosodic context specifically, Steffman (2019b) also showed that variation in context can indeed up-weight these effects in terms of their magnitude. In the present case, this account would predict increased listener attention to the (more variable) target in Experiment 2, as compared to Experiment 1, and thus the up-weighting of intrinsic cues (VOT and vowel duration) and down-weighting of contextual cues (phrasing). This account, however, is speculative, and exploring how attention mediates the effects of phrasing and other similar effects seems to be a promising further research direction.

The results of Experiment 2, in line with our interpretation of Experiment 1's results, show that listeners need vowel duration to be even longer to cue a fortis stop when AP-initial, in line with AP-initial vowel lengthening associated with the fortis stop. This suggests that listeners' expectations of vowel duration shift based on AP phrasing. Experiment 2's results thus confirm that vowel duration is used by listeners as a cue to the fortis-aspirated contrast in our stimuli, and importantly, that listeners' categorization of vowel duration itself shifts as a function of prosodic phrasing (at fixed or pooled VOT values).

5. General discussion

In two experiments we tested how variation in phrasing at the level of the AP in Seoul Korean exerts a mediating influence on listeners' perception of the fortis-aspirated stop contrast, cued by VOT and vowel duration. The experiments together show that a prosodic structure, which reflects the way in which an utterance is phrased, can modulate how these cues are perceived. This is notable given the care we took to design our stimuli to avoid the possible confound driven by variation in temporal context. As described in Section 1.1, this question has been a persistent one in recent studies (Kim and Cho, 2013; Mitterer et al., 2016; Steffman, 2019a). The results presented here allow us to show that, independent of variation in temporal context (adjacent segmental durations), listeners are influenced by prosodic phrasing as they map durational cues to segmental categories. More broadly, the results are consistent with the view that speech perception is modulated in reference to higher-order prosodic structure, regardless of whether phonetic cues used by listeners to compute prosodic structure are temporal or f0-based.

Similar claims about the importance of prosodic context in segmental perception have recently been made in the literature, with evidence coming from perception of cues which are not durational, e.g., glottalization in Maltese (Mitterer et al., 2019) and formant cues to vowel quality in American English (Steffman, 2021a,b). These findings, and the present ones, can be taken to offer general support for a model of language processing which involves both segmental and prosodic parsing in word recognition. As noted in section 2.1, the *prosodic analysis* model holds that prosodic information is integrated at a post-lexical stage in processing, as a mediating factor in lexical competition (Cho et al., 2007; Kim et al., 2018b; Mitterer et al., 2019; McQueen and Dilley, 2020; Steffman, 2020). The model thus predicts a relatively delayed influence of prosodic boundary information, and eyetracking studies have corroborated this idea, showing a clear temporal delay of the use of prosodic boundaries in online processing (Kim et al., 2018b; Mitterer et al., 2019). In our case, the findings may implicate a similar sort of processing, predicting that we should see a relatively delayed impact of the phrasing manipulation (delayed, for example, in relation

to the uptake of VOT as a cue to the stop contrast). Such a delayed timecourse would be clearly different from the near-immediate impact of temporal and spectral context in online processing (e.g., Reinisch and Sjerps, 2013; Toscano and McMurray, 2015). This seems to be a promising further test of the claims we forward here. If found, this would corroborate the idea that listeners process the available acoustic-phonetic input, whether spectral or temporal, to compute an abstract prosodic structure in parallel with segmental analysis, and that they use the integrated information relatively later in speech processing to inform (calibrate/modulate) lexical access/competition, as described in the prosodic analysis model.

On the other hand, some prosodic contextual influences, such as prominence-related modulation, have been documented to occur relatively rapidly in online processing. For example, Steffman (2020, 2021b) found that listeners rapidly integrate the relative prominence of a word with their perception of vowel contrasts, though when prominence is conveyed by phrasal modulations in duration and pitch,[12] the effect is slow-growing and is strengthened later in the timecourse of processing. In comparison, Steffman (2020) found that a strictly localized cue to prominence, vowel-initial glottalization (Dilley et al., 1996; Garellek, 2014), showed a near-immediate influence in vowel processing that peaked early in time. This was in clear contrast to the timecourse for prominence cued by phrasally distributed variations in pitch and duration.

This asymmetry between local and temporally more extended prosody-signaling contexts motivated the proposal that, at least in the case of prosodic prominence, multiple stages of processing are implicated: both an immediate adjustment for prosodic information conveyed by localized phonetic context (i.e. glottalization at vowel onset), and a later-stage integration of a parsed prosodic structural representation that encodes (phonological)

[12]This phrasal prominence manipulation signaled the presence or absence of focus on a word preceding the target word, (the word "say" in the sentence "I'll say [the target] now"). In the case that the target was preceded by focus it was non-prominent (post-focal), as compared to a broad focus context in which it was relatively prominent.

prominence. This proposal, which Steffman (2020) called Multi-stage Assessment of Prominence in Processing (MAPP), is a hypothesis which complements the prosodic analysis model in postulating that certain contextual prosodic information can impact processing without delay. In this light, exploring what *types* of contexts are processed earlier or later becomes an interesting research question. In the case of the present study, because we are dealing with a small phrasal domain, an earlier influence in online processing could index the computation of more local cues which are built incrementally and later integrated into a larger phrasal structure, as described by Steffman (2020) (though we note here that Kim et al., 2018b found a delayed timecourse in phonological inferencing based on AP phrasing). In summary, insights from the MAPP proposal lead us to consider that some prosodic contextual influences (especially when they stem from prominence-related prosodic structure) might begin relatively early in processing, and strengthen as the listener accumulates more information about the prosodic structure of an utterance (see McQueen and Dilley 2020 for a Bayesian framing of this sort of process). Future work testing how the timecourse of processing correlates with the size of prosodic domains (where we might predict smaller domains are processed more quickly) and the locality of the relevant prosody-signaling cues to a given target segment will thus help enrich the prosodic analysis model and related models (e.g., McQueen and Dilley, 2020). This future work will also address if the insights gleaned from prosodic prominence in American English, which motivated the MAPP proposal, extend to other languages and prosodic features, as a further test for this hypothesis. As is clear from the forgoing discussion, much of the research which has examined questions of timing in processing in this domain has focused on eye-movement and reaction time data (Cho et al., 2007; Kim et al., 2018b; Mitterer et al., 2019; Steffman, 2020, 2021b). As noted by McQueen and Dilley (2020, p 520) "[...] much work still needs to be done to specify the brain mechanisms that support spoken-word recognition as a process that depends on parallel inferences about segmental content and prosodic structures". It has been noted in several places (Mitterer et al., 2019; McQueen and Dilley, 2020) that work focusing on neural entrainment (Ding et al., 2016; Giraud and Poeppel, 2012) may offer a useful avenue for

understanding how neural oscillations at different time scales support the understanding of segmental and prosodic information. We highlight this here as an area in need of future work, which will help us relate current findings to their neural signatures and the brain structures which are implicated.

We also found that certain cues appear to be more sensitive to prosodic context than others, where in our case, listeners shift categorization in line with how vowel duration, but not VOT, is modified by patterns of AP-initial strengthening. To our knowledge, this is the first study comparing perception of multiple prosodically-modulated cues in this light, and the apparent importance of vowel duration seen here merits exploration in future work. To the extent that a vowel serves as a better cue to prosodic structure as compared to VOT (i.e. being sonorous, and bearing tonal information), we might predict that computation of prosodic structure would be more tightly linked to this cue. That is, while both VOT and vowel duration co-vary with prosodic phrasing, the link between vowels and prosodic structural parsing may be tighter. This will need to be tested in other cases where two cues undergo prosodic modulation and make different predictions about perceptual responses.

Alternatively, the down-weighting of VOT as a prosodically-modulated cue could relate to an ongoing sound change in the Seoul dialect of Korean. VOT in aspirated stops (though still robustly longer than fortis stops) has become shorter for younger speakers (Kang, 2014; Choi et al., 2020), so that VOT no longer serves to distinguish between aspirated and *lenis* stops in many contexts, though this is intricately linked with prosodic structure as described by Choi et al. (2020). For example, VOT still distinguishes aspirated from lenis stops in non-prominent, phrase-medial contexts. The contrast between aspirated and lenis stops is additionally, and robustly, cued by pitch. As VOT becomes a redundant cue, particularly for phrase-initial aspirated stops (in relation to lenis stops), this may entail a prosodic-structurally conditioned re-weighting, such that listeners are less attuned to prosodically-driven variation in VOT duration, thus not shifting categorization in line with initial strengthening (lengthening) of VOT.

One limitation of the present study is the examination of only two cues, both durational

in nature. As noted above, fortis and aspirated stops also vary in their effects on voice quality spectral measures (e.g., Cho et al., 2002). Future work on this particular contrast will thus benefit from including other cue variations, particularly in light of the now-established effect of phrasing on the perception of durational cues. The present study also employed a contextual manipulation that was categorical and had only two possible levels (AP-initial vs. medial). Another useful future direction may therefore be the implementation of contextual cues which are more gradient in nature, for example, an f0 continuum which ranges from a clear AP boundary endpoint to a lack-of-boundary endpoint. Examining how listeners integrate continuous contextual information with perception of segments, and if perceptual adjustments vary linearly or categorically with a context continuum, would thus inform how prosodic-structural context is being represented and parsed by listeners. This may be fruitfully framed in light of Bayesian accounts of prosodic parsing, as in the Bayesian Prosody Recognizer (McQueen and Dilley, 2020), where signal-based likelihoods in favor of a particular prosodic parse are computed in parallel to a segmental characterization of the signal.

Another promising future direction could link the present results to what is already known about the usefulness of the Korean AP, and prosodic phrasing more generally, in word segmentation (Christophe et al., 2004; Kim and Cho, 2009; Tremblay et al., 2012) and phonological inferencing (Kim et al., 2018b). Taking the present results in hand with this other literature, it is clear that small phrasal domains play an important role at multiple levels of processing. Future work may benefit from testing how these domains interact. For example, if AP-initial lengthening of a post-stop vowel helps signal a word boundary, to what extent is this dependent on the segmental identity of the stop? In other words, should a post-fortis vowel be longer than a post-aspirated vowel to help listeners parse a word boundary? Put more generally: to what extent does the perception of segmentation cues (including phrasing) relate to segmental cues, and relate to the expectation of how a segmental contrast should be realized based on phrasing? Parallel processing, as conceived of in current models (Cho et al., 2007; McQueen and Dilley, 2020; Steffman, 2020) certainly

allow for this sort of bidirectional influence, though the extent to which segmental parsing influences prosodic parsing, and the intricate interactions between these domains remain to be explored empirically and fit into these models (cf. Mitterer et al., 2021). Thus, testing how AP-initial durational cues impact perception of phrasing (facilitating word segmentation) would essentially ask a reversal of the question addressed in this study, and would help us better understand how prosodic parsing is impacted by segmental information. Further exploring questions such as these will help us expand on the present results and inform the development of a theory of prosodic parsing in speech recognition.

Acknowledgments

We thank Hyunjung Joo, Jungyun Seo, and Seungwoo Baek for their help with data acquisition, and the participants in the experiments. This work was supported in part by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2018S1A5A2A03036736) awarded to author SK.

Appendix

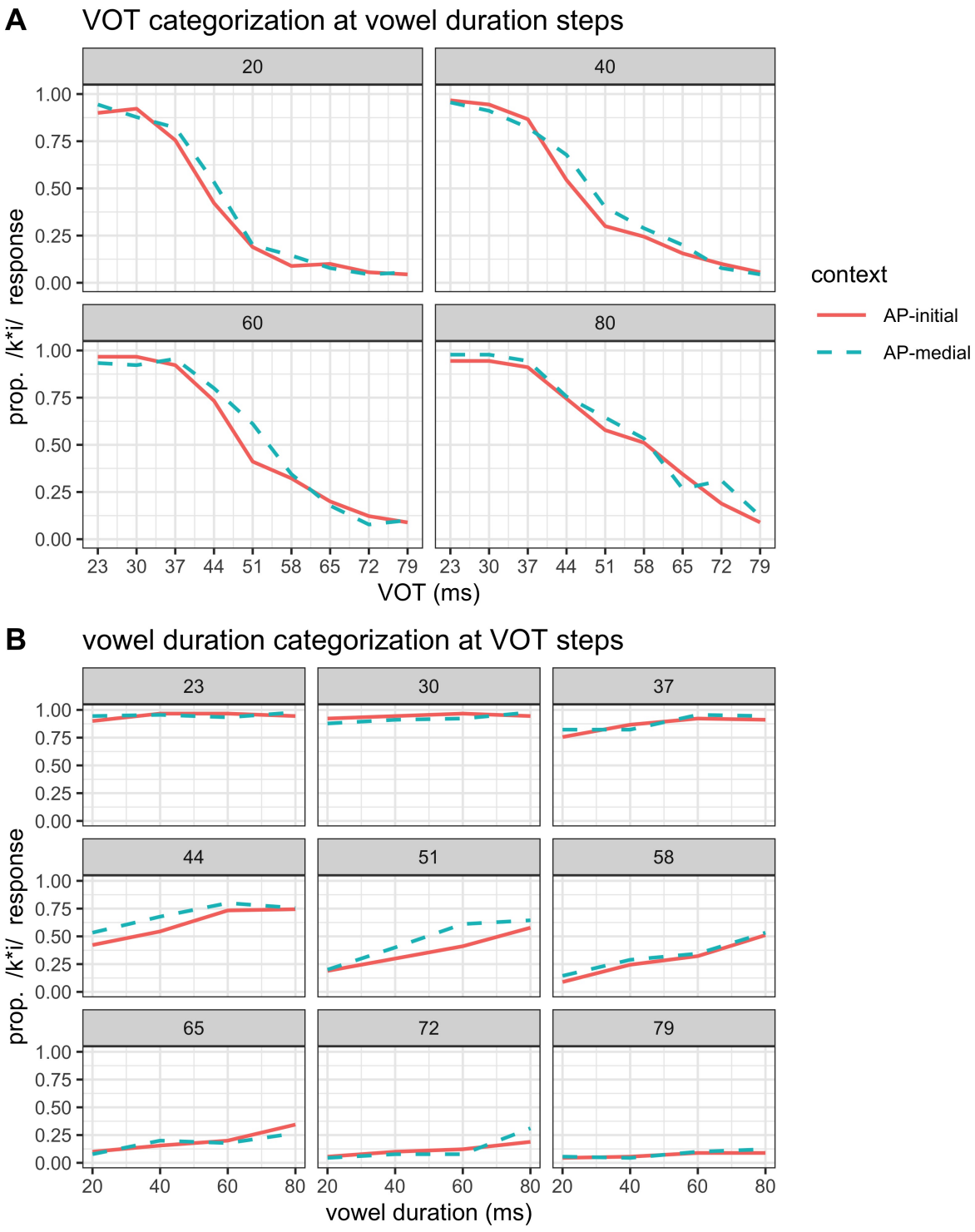

Figure 6: Categorization of VOT at vowel duration steps, where each panel is a step (A), and of vowel duration at VOT steps (B), split by context in Experiment 2.

References

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.

Bosker, H. R., Sjerps, M. J., and Reinisch, E. (2020). Temporal contrast effects in human speech perception are immune to selective attention. *Scientific reports*, 10(1):1–11.

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

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

Cho, T. (1996). *The role of variable vowel duration in differentiating stop phonation in Korean*. PhD thesis, Acoustical Society of America.

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 Redford, M. A., editor, *The Handbook of Speech Production*, pages 505–529. John Wiley & Sons, Inc.

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

Cho, T. and Jun, S.-A. (2000). Domain-initial strengthening as enhancement of laryngeal features: Aerodynamic evidence from Korean. *UCLA working papers in phonetics*, pages 57–70.

Cho, T., Jun, S.-A., and Ladefoged, P. (2002). Acoustic and aerodynamic correlates of korean stops and fricatives. *Journal of Phonetics*, 30(2):193–228.

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

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.

Choi, H. (2011a). Interface between speech production and perception: Analysis of stop and vowel duration. *Korean Journal of Linguistics*, 36(3):815–842.

Choi, H. (2011b). Vowel duration as a perceptual cue for preceding stop laryngeal contrast in Korean. In *ICPhS*, volume 17, pages 17–21.

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

Choi, S. and Jun, J. (1998). Are korean fortis and aspirated consonants geminates? *Language Research*, 34(3):521–546.

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.

Chung, H., Kim, K., and Huckvale, M. (1999). Consonantal and prosodic influences on korean vowel duration. In *Proceedings of Eurospeech*, volume 99. Citeseer.

Diehl, R. L. and 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., and Ostendorf, M. (1996). Glottalization of word-initial vowels as a function of prosodic structure. *Journal of Phonetics*, 24(4):423–444.

Ding, N., Melloni, L., Zhang, H., Tian, X., and Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature neuroscience*, 19(1):158–164.

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

Giraud, A.-L. and Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. *Nature neuroscience*, 15(4):511–517.

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

Jun, S.-A. (1993). *The phonetics and phonology of Korean prosody. Ohio State University*. PhD thesis, Thesis/Dissertation.

Jun, S.-A. (1995). Asymmetrical prosodic effects on the laryngeal gesture in Korean. papers in laboratory phonology iv: phonology and phonetic evidence, ed. by B. Connell and A. Arvaniti.

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*, pages 189–226.

Jun, S.-A. (2000). K-tobi (Korean tobi) labelling conventions. *Speech Sciences*, 7(1):143–170.

Jun, S.-A., editor (2005). *Prosodic typology: The phonology of intonation and phrasing*, volume 1. Oxford University Press.

Jun, S.-A., editor (2014). *Prosodic Typology II: The phonology of intonation and phrasing*. Oxford University Press.

Jun, S.-A. and Cha, J. (2015). High-toned [il] in Korean: Phonetics, intonational phonology, and sound change. *Journal of Phonetics*, 51:93–108.

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

Keating, P. (2006). Phonetic encoding of prosodic structure. *Speech production: Models, phonetic processes, and techniques*, pages 167–186.

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

Kim, D.-W. (2002). The vowel length as a function of the articulatory force of the following consonants in korean. *Speech Sciences*, 9(3):143–153.

Kim, S. (2004). *The role of prosodic phrasing in Korean word segmentation*. PhD thesis, UCLA.

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

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

Kim, S., Cho, T., and McQueen, J. M. (2012). Phonetic richness can outweigh prosodically-driven phonological knowledge when learning words in an artificial language. *Journal of Phonetics*, 40(3):443–452.

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.

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).

Koo, H. S. (1986). *An experimental acoustic study of the phonetics of intonation in standard Korean*. University of Texas at Austin.

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

Lenth, R., Singmann, H., Love, J., Buerkner, P., and Herve, M. (2018). emmeans: Estimated Marginal Means, aka Least-Squares Means.

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

Makowski, D., Ben-Shachar, M. S., Chen, S., and Lüdecke, D. (2019a). Indices of effect existence and significance in the bayesian framework. *Frontiers in psychology*, 10:2767.

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

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.

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

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.

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

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.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of experimental psychology: General*, 115(1):39.

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.

RStudio Team (2020). *RStudio: Integrated Development Environment for R v.1.3.959*. RStudio, PBC, Boston, MA.

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. (2020). *Prosodic prominence in vowel perception and spoken language processing*. PhD thesis, University of California, Los Angeles.

Steffman, J. (2021a). Contextual prominence in vowel perception: Testing listener sensitivity to sonority expansion and hyperarticulation. *JASA Express Letters*, 1(4):045203.

Steffman, J. (2021b). Prosodic prominence effects in the processing of spectral cues. *Language, Cognition and Neuroscience*, pages 1–26.

Tehrani, H. (2020). Appsobabble: Online applications platform. `https://www.appsobabble.com`.

Toscano, J. C. and 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.

Tremblay, A., Coughlin, C. E., Bahler, C., and Gaillard, S. (2012). Differential contribution of prosodic cues in the native and non-native segmentation of French speech. *Laboratory Phonology*, 3(2):385–423.

Wade, T. and Holt, L. L. (2005). Perceptual effects of preceding nonspeech rate on temporal properties of speech categories. *Perception & psychophysics*, 67(6):939–950.