JASA ARTICLE

Tonal cues to prosodic structure in rate-dependent speech perception

Jeremy Steffman[1,a)] and Sun-Ah Jun[2]
[1]Deparment of Linguistics, Northwestern University, Evanston, Illinois 60201, USA
[2]Deparment of Linguistics, University of California Los Angeles, Los Angeles, California 90095, USA

ABSTRACT:
This study explores how listeners integrate tonal cues to prosodic structure with their perception of local speech rate and consequent interpretation of durational cues. In three experiments, we manipulate the pitch and duration of speech segments immediately preceding a target sound along a vowel duration continuum (cueing coda stop voicing), testing how listeners' categorization of vowel duration shifts based on temporal and tonal context. We find that listeners perceive the presence of a phrasal boundary tone on a lengthened syllable as signaling a slowdown in speech rate, shifting perception of vowel duration, with effects that are additive when crossed in a 2×2 (pitch $\times$ duration) design. However, an asymmetrical effect of pitch and duration is found in an explicit duration judgement task in which listeners judge how long a pre-target syllable sounds to them. In explicit rate judgement, only durational information is consequential, unlike the categorization task, suggesting that integration of tonal and durational prosodic cues in rate-dependent perception is limited to implicit processing of speech rate. Results are discussed in terms of linguistic information in rate-dependent speech processing, the integration of prosodic cues, and implicit and explicit rate processing tasks. © *2021 Acoustical Society of America*.
https://doi.org/10.1121/10.0007222
(Received 24 January 2021; revised 20 October 2021; accepted 22 October 2021; published online 19 November 2021)
[Editor: Melissa Michaud Baese-Berk] Pages: 3825–3837

I. INTRODUCTION

A. Background

Speech rate varies within and across utterances, speakers, and languages [e.g., Miller *et al.* (1984), Pellegrino *et al.* (2011), and Quené (2008), (2013)]. Variability in speech rate creates variability in the temporal structure of speech, modulating the distribution of acoustic information in the speech signal over time. Accordingly, one well-established finding in the speech perception literature is that listeners' interpretation of durational cues in speech varies as a function of rate [e.g., Miller and Volaitis (1989) and Wade and Holt (2005)]. This sort of rate-sensitive perceptual adjustment is important for language comprehension given the wide variety of phonological contrasts in language that are cued in part by durational information [e.g., voice onset time (VOT), contrastive vowel length, etc.]. However, interestingly, perception of speech rate is not entirely veridical, and is mediated by various linguistic and cognitive factors [e.g., Bosker and Reinisch (2017), Bosker *et al.* (2017), Reinisch (2016), and Steffman (2019)]. The present study addresses one such factor: the *prosody* of a language, which systematically organizes the temporal and tonal structure of speech. In three experiments we explore how listeners' interpretation of durational cues is mediated by prosodic patterning of duration and pitch, building on recent research.

Speech rate effects, as we will call them, refer to observed adjustments in categorization and processing of temporal cues in speech as a function of speech rate. In this study we are primarily concerned with *proximal* speech rate effects. In this context, "proximal" refers to temporally localized or adjacent (in terms of e.g., syllables, segments) to a given target sound (as compared to distal rate). Various findings suggest that speech rate effects in general can be described in terms of a general auditory mechanism that operates early in processing: they are observed with non-speech precursors (Wade and Holt, 2005), in nonhuman species (Welch *et al.*, 2009), are immune to modulation of attention and cognitive load (Bosker *et al.*, 2017, Green *et al.*, 1997), and play out rapidly in online processing (Reinisch and Sjerps, 2013). To account for proximal rate effects, Diehl and Walsh (1989) proposed the mechanism of *durational contrast*, wherein the "perceived length of a given acoustic segment is affected contrastively by the duration of adjacent segments" (p. 2154). Consider an example: formant transition duration into a vowel is one temporal cue distinguishing the stop /b/ from the approximant /w/, where /w/ shows a longer transition duration. Wade and Holt (2005) created a target syllable that varied in transition duration, categorized as /bɑ/ or /wɑ/. When a longer acoustic event (in comparison to a shorter one) preceded a target syllable, listeners required overall longer transition duration to perceive /w/. In other words, transition duration was perceived as relatively short (/b/-like) in relation to preceding context, explainable as resulting from the perceptual *contrast* between transition

a)Electronic mail: jeremy.steffman@northwestern.edu, ORCID: 0000-0003-3675-910X.

duration and preceding material. This mechanism of durational contrast accounts for a large body of findings in the literature [see, e.g., Bosker (2017) for discussion].

In addition to adjacent acoustic segment durations, another influence on listeners' processing of durational cues is the integration of other acoustic dimensions with perception of duration. These effects, which will be referred to as *psychoacoustic effects* here, are evidenced when one dimension of an auditory stimulus influences perception of another [e.g., Turk and Sawusch (1996) and Prince (2011)]. Most relevant to the present study is the finding that pitch influences perceived duration [e.g., Brigner (1988) and Simko *et al.* (2016)]. Higher pitch is perceived as longer, and more dynamic pitch contours are likewise perceived as longer. Perceived duration as a function of changing pitch also shapes listeners' interpretation of durational cues to linguistic contrasts, such as vowel duration as a cue to obstruent voicing (Steffman and Jun, 2019).

Durational contrast and psychoacoustic effects of pitch on perceived duration thus both constitute two influences of listeners' processing of temporal cues in speech. However, in addition to these purely signal-based auditory effects, a variety of other factors shape listeners' perception of speech rate and resultant processing of durational cues (Bosker *et al.*, 2017; Bosker and Reinisch, 2017). For example, fast speech processes related to the phonology of a language, like reduction, lenition, etc., make speech sound faster to listeners in comparison to speech that has the same rate (calculated in e.g., the number of syllables or segments per unit of time) but lacks fast speech processes (Reinisch, 2016). Reinisch (2016) found that fast speech processes (independent of actual speech rate) accordingly shifted listeners' perception of a durational contrast (German vowel length), suggesting that calculation of speech rate can incorporate multiple cues, some of which relate to linguistic patterns (Bosker and Reinisch, 2017; Reinisch, 2016; Steffman, 2019; Toscano and McMurray, 2015). Adding further nuance, Pitt *et al.* (2016) tested how changes in contextual speech rate impacted listeners' perception of durational cues signaling a reduced function word (e.g., "minor or child" versus "minor child," where perception of a reduced "or" is rate-dependent). Rate effects were observed only when preceding speech material was intelligible to listeners: when unintelligible filtered speech or pure tones, both of which still conveyed the relevant timing information, preceded the critical portion of speech, the effect disappeared [cf. Wade and Holt (2005), among others, successfully used non-speech precursors to cue rate]. Pitt *et al.* thus show that durational contrast effects are not always observed when they might be expected, in accordance with other research which has shown that proximal contrast effects can disappear when there is orthogonal variation in overall rate (Bosker, 2017), and in the presence of durational rhythmic alternations (Kidd, 1989; Steffman, 2021a).[1] The incorporation of other cues in rate-dependent perception, and the occasional absence of durational contrast effects, suggest some counterevidence to the notion that a domain-general contrast mechanism is solely responsible for speech rate related perceptual adjustments.

Given these findings, the sorts of cues that listeners utilize in rate-dependent perception, and the extent to which durational contrast effects are relevant in various contexts, are somewhat open questions. A pertinent line of inquiry concerns the possible mediating role of linguistic factors, though relatively little is known about how perception of more localized changes in rate may be influenced by linguistic factors (Mitterer *et al.*, 2016). One aspect of linguistic structure that plays a crucial role in organizing localized changes in speech rate is *prosody*. The prosodic organization of a language introduces systematic patterning in temporal patterns [e.g., Cho (2015), (2016), Turk and Sawusch (1997), and Turk and Shattuck-Hufnagel (2007)], and accordingly can be conceptualized as structure that influences "… the domain and distribution of durational effects" [Turk and White (1999), p. 171].

We focus on just one prosodically driven temporal pattern: pre-boundary (or phrase-final) lengthening in American English. This refers to the temporal expansion of linguistic units preceding a phrasal boundary, usually the boundary of an intonational phrase [e.g., Cho (2015), Edwards *et al.* (1991), and Turk and Shattuck-Hufnagel (2007)]. Pre-boundary lengthening is largely localized to the phrase-final rhyme (Wightman *et al.*, 1992), though in polysyllabic words it can spread further leftwards to the final syllable onset and pre-final syllables that bear stress (Turk and Shattuck-Hufnagel, 2007). Temporal modulations associated with phrase-final lengthening also co-occur with a suite of acoustic cues to phrasal boundaries, including changes in voice quality (Epstein, 2002; Redi and Shattuck-Hufnagel, 2001) and pitch (Liberman and Pierrehumbert, 1984).

In the intonational phonology of Mainstream American English (Beckman and Pierrehumbert, 1986; Ladd, 2008; Pierrehumbert, 1980), one function of the tonal melody of an utterance is to cue the end of phrasal constituents via boundary tones. Boundary tones at the right edge of an intonational phrase (IP) can take various shapes, modeled as the composition of low (L) and high (H) tonal targets. For example, a typical declarative sentence ends in falling pitch, labeled L-L%, where L- refers to the tone demarcating the boundary of a smaller prosodic constituent, the intermediate phrase (ip), and L% refers to the boundary tone of the IP. In contrast, a high rising boundary tone, labeled H-H%, is typically used in yes-no questions. These tonal melodies, as a property of their boundary marking function, co-occur with pre-boundary lengthening at the right edge of an IP.

How should intonational pitch patterns factor into listeners' perception of speech rate? It is well known that pitch-based cues to rhythmic patterning influence listeners' perception of rhythmic grouping [e.g., Dilley and McAuley (2008) and Morrill *et al.* (2014)], though the role of pitch in rate-dependent perception is relatively unexplored. Boundary tones present a clear case of a pitch-based cue that co-occurs with systematic changes in the temporal structure of speech. We can conceptualize pre-boundary

lengthening as a localized change in segment duration, i.e., a proximal slow-down in speech rate. In this light, boundary tones could constitute a possible cue to such a slow-down, on the basis of their patterned co-occurrence. We can therefore ask, do boundary tones shape listeners' perception of speech rate changes?

Though little previous research on this topic exists, one study suggests that listeners do indeed structure their perception of rate in accordance with intonational patterns. Steffman (2019) tested how a low-rising (L-H%) boundary tone, often used to cue a continuation rise, influences perception of speech rate. In a two-alternative forced choice task, listeners categorized a voice onset time (VOT) continuum as /p/ or /b/. VOT, the duration between the release of an oral stop closure and the onset of voicing in a subsequent vowel, is a robust rate-dependent cue to voicing in languages like American English. Longer VOT is perceived as voiceless /p/ and shorter VOT is perceived as voiced /b/ [e.g., Miller and Volaitis (1989) and Lisker and Abramson (1970)]. The target word was placed in a carrier phrase "I'll say __ again," with pitch accents on the word "I'll" and the target word. The crucial manipulation was the f0 and duration of the pre-target syllable, "say," which crossed the length of the syllable (short versus long) with the pitch over that syllable, which bore either a low-rising or high flat contour. In English, a low-rising (i.e., bitonal L-H%) boundary tone can occur on a single IP-final syllable (which is lengthened due to phrase final lengthening). The short condition with the low-rising L-H% f0 pattern accordingly takes a boundary-marking tonal pattern and compresses it onto a single syllable which lacks phrase-final lengthening (a shorter-than-usual temporal interval for the tonal pattern).[2] Steffman found that categorization of subsequent VOT was affected by this f0 manipulation such that in the short condition only, where the L-H% f0 pattern was compressed, listeners *increased* /p/ responses (relative to the flat pitch condition), suggesting that a tonal pattern that is co-occurrent with phrase-final lengthening was perceived as an *increase in speech rate* when compressed, i.e., when two tonal targets are realized over a shorter temporal interval than is typical [see Steffman (2019) for details].

This offers suggestive evidence for the importance of tonal cues to prosodic structure in listeners' perception of the temporal structure of speech. However, many questions remain. How do listeners use tonal cues to prosodic boundaries in speech rate perception under more typical circumstances (i.e., when a boundary tone is not compressed)? Do these effects relate to psychoacoustic processing of pitch and duration? The present study addresses these outstanding questions.

B. The present study

As outlined above, the central role of prosody in organizing the temporal structure of speech motivates us to test how listeners' perception of speech rate may be influenced by prosodic cues. We examine how tonal cues to prosodic structure influence perception of speech rate, testing perception of a durational cue: vowel duration as a cue to coda obstruent voicing. In American English, vowels are longer preceding voiced obstruents [e.g., Chen (1970), Moreton (2004), Summers (1987), and Wolf (1978)], and this is a reliable cue to voicing for listeners [e.g., Raphael (1972) and Steffman and Jun (2019)]. As would be expected, perception of this durational cue is influenced by contextual speech rate [e.g., Heffner *et al.* (2017) and Steffman (2019)], such that it presents a useful test case for our question.

As mentioned earlier, Steffman found that a compressed boundary tone was interpreted by listeners as an increase in speech rate. However, this is not the normal state of affairs in spoken language—boundary tones in spoken language co-occur with lengthening. In the present study we effectively ask the opposite of the question addressed in Steffman (2019): we test if tonal cues to a prosodic boundary influence listeners to perceive a *slow down* in speech rate. In other words, does the reliable co-occurrence of boundary tone cues with pre-boundary lengthening (i.e., local rate slowing) shape listeners' perception of speech rate?

Steffman (2019) further manipulated pre-target duration by cross-splicing material from IP-final position, such that the condition in which pre-target duration was long also included falling intensity and changes in voice quality, in combination with pitch as a boundary cue. In the present study, we control our stimuli to vary in *only* pitch and duration, allowing us to test how pitch and lengthened duration alone are interpreted by listeners. The tonal cue we selected was a low-falling (L-L%) IP-final boundary tone, the most frequently used in American English [Dainora, 2001, 2006]. By manipulating the presence/absence of this boundary tone, we can test how pitch-based cues to boundary may influence perception of rate, and more generally therefore how rate-dependent perception integrates multiple sources of information about speech rate. Addressing these questions will help us better understand how rate-dependent cues in speech are processed, and how prosody influences speech perception and word recognition more generally [Mitterer *et al.*, 2016; Mitterer *et al.*, 2019; Steffman and Katsuda, 2020]. The structure of the paper is outlined after describing the stimuli in Sec. II C below.

C. Materials

The materials used in all experiments were created by resynthesizing the speech of a ToBI-trained male speaker of American English. Speech material was recorded using an SM10A Shure™ microphone and headset, at 44.1 kHz (32 bit). Recording was carried out in a sound-attenuated booth. The target sound itself was drawn from a vowel duration continuum, manipulated to vary only in duration, which listeners categorized as one of two English words: "coat" or "code." These two words were selected because they are relatively matched in lexical frequency ("code"

$Log_{10WF} = 3.43$; "coat" $Log_{10WF} = 3.33$), as calculated from the $SUBTLEX_{US}$ corpus (Brysbaert and New, 2009). The basis for stimulus creation was two utterances, shown in (1) and (2) with ToBI labels below.

(1) I'll say code now

H* H* L-L%

(2) I'll say code now

H* L-L% H* L-L%

In (1), the target word "code" is produced in the middle of a single IP, with relatively high pitch, and short duration of the pre-target syllable, "say." In (2), the target is preceded by an IP boundary, marked by lengthened duration (i.e., local speech rate slowing) and a low-falling boundary tone on the pre-target syllable.

The target sound, produced as "code," was excised from (1). This token, which served as the base for the creation of the continuum, had a vowel duration of approximately 120 ms. Duration was manipulated using PSOLA (Moulines and Charpentier, 1990), as implemented in PRAAT (Boersma and Weenik, 2020). First, audible voicing during the beginning of the stop closure was spliced out so that the closure for the stop was totally silent. This was done to render stop-internal cues to voicing ambiguous such that only the duration of the vowel cued the contrast. Duration of the vowel was then manipulated to range from 60 to 150 ms in 15 ms steps, for seven steps including the endpoints. The target vowel continuum was judged by the authors and one additional speaker of American English to sound clearly like "coat" when vowel duration was short and "code" when vowel duration was long. Each target was then spliced into four carrier phrases which differed in pre-target duration and f0, described below.[3]

The goal in creating four carrier phrases used in the present study was to cross pre-target pitch and duration (with two values for each parameter), with all of these manipulations occurring on the vowel immediately preceding the target word. The starting point for manipulations was the vowel [eɪ] in "say" as produced in (1), which was approximately 125 ms in duration. To create the two duration conditions (SHORT and LONG) used in experiments 1 and 2, the vowel was resynthesized to have a duration of 125 and 200 ms, which was the approximate duration of the pre-boundary vowel in (2). We subsequently adjusted the duration of this pre-target vowel in experiment 3, which will be described in Sec. IV A. The durational manipulation was crossed with a f0 manipulation, whereby the f0 on the target vowel was resynthesized in two conditions. One condition used f0 from the pre-target vowel as produced in (2). This was realized as an L-L% boundary tone and was accordingly relatively low and falling (onset: 115 Hz; offset: 100 Hz). This pitch condition will be referred to as the BOUNDARY TONE condition, abbreviated BT. The other pitch condition was created by overlaying the pitch from another production of (1).[4] This pitch pattern was slightly dipping over "say," realizing a sag between the two adjacent H* pitch targets. This sagging pitch was relatively high and less falling as compared to the boundary tone condition (onset: 125 Hz; offset: 112 Hz). These pitch contours were overlaid on both LONG and SHORT conditions, creating a 2×2 crossing of pre-target duration and pitch. All four conditions are shown in Fig. 1.

With the aforementioned structure of the stimuli, we will address the questions outlined in Sec. II B in three experiments. Experiment 1 will test how duration only, pitch only, and both cues combined influence categorization of the target continuum in three sub-experiments. Experiment 2 will examine how these effects translate into an explicit durational judgement task in which listeners evaluate the duration of the stimuli. Experiment 3 will examine how the effects in experiment 1 play out in a 2×2 design, while also enhancing the magnitude of the duration manipulation (described in Sec. IV A).

II. EXPERIMENT 1

Experiment 1 was a forced-choice task in which participants categorized the target continuum. It consisted of three

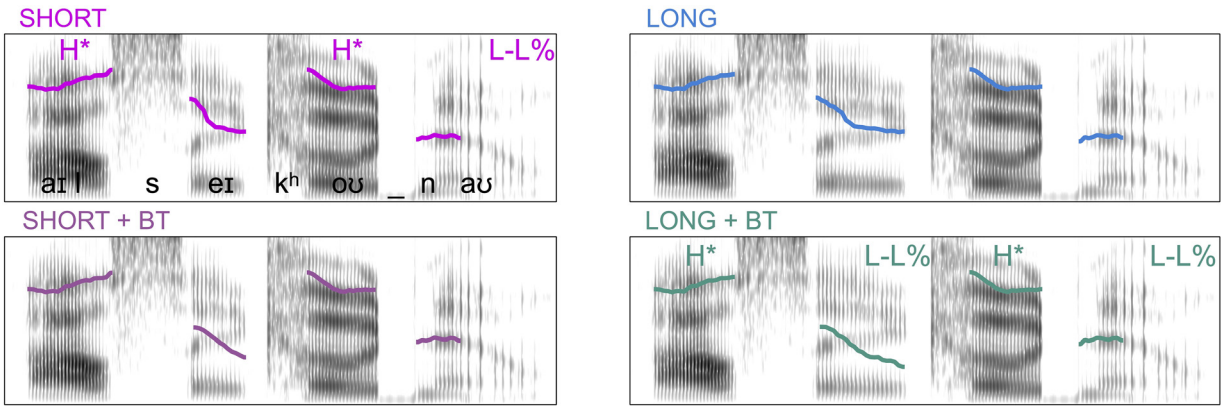

FIG. 1. (Color online) Spectrograms with overlaid pitch tracks showing the conditions used in all experiments. A segmental transcription is given in the SHORT panel of the Figure, with ToBI labels for hypothesized listener interpretation given below the SHORT and LONG +BT condition. The pitch range for the pitch track is 85–150 Hz. The frequency range for the spectrogram is 0-5000 kHz. Note the phrase-final "now" shows creaky phonation such that pitch is not tracked.

https://doi.org/10.1121/10.0007222

sub-experiments, each of which compared two of the pre-target duration/pitch conditions. In experiment 1a, we compared the SHORT condition to the LONG condition, such that only pre-target duration was manipulated. In experiment 1b, we compared the SHORT condition to the SHORT+BT condition, such that only pitch varied. In experiment 1c, we compared the SHORT condition to the LONG+BT condition in which both pre-target duration and pitch were manipulated. In experiment 1c, with both duration and pitch cueing a boundary, we predicted to see larger or more robust speech rate effects, as compared to the duration-only comparison in experiment 1a. Experiment 1b served a test for an independent effect of pitch, in light of the possible influence of pitch on perceived duration, or potentially as an independent cue to rate changes.[5]

A. Participants

All participants were self-reported native speakers of American English with normal hearing. Participants were students at UCLA, who received course credit for their participation. Different participants were recruited for each sub-experiment. 36 participants were recruited for experiment 1a, 43 participants for experiment 1b, and 40 for experiment 1c.

B. Procedure

Participants completed testing in a sound-attenuated room, seated in front of a desktop computer. Stimuli were presented binaurally via a Peltor^{TM} 3M^{TM} listen-only headset, with the volume adjusted to a comfortable listening level. Before testing, participants were told they would listen to a native English speaker saying "I'll say x now," and their task was to decide if the speaker said the word "coat" or "code" for "x." During testing, participants heard a stimulus and were presented visually with the texts "coat" and "code" centered in either half of the computer screen. Participants indicated their choice of word via a key press on the computer keyboard, where an "f" key press indicated the left-side choice, and a "j" key press indicated the right-side choice. The side of the screen on which "coat" and "code" appeared was counterbalanced across participants in all experiments. Prior to testing in each sub-experiment, participants completed four practice trials in which they heard the continuum endpoints in both pre-target prosodic conditions. After these practice trials, the test trials began. In the test trials participants categorized 16 repetitions of each of the 14 unique stimuli (7 steps in the two pre-target conditions), for a total of 224 trials in each sub-experiment. Stimuli were completely randomized, and split into two blocks, with a short self-paced break given halfway through the test trials. The experimental procedure for each sub-experiment took approximately 20 min to complete.

C. Results and discussion

Results were assessed separately for each sub-experiment by a Bayesian mixed-effects logistic regression, implemented using the *brms* package in R (Bürkner, 2017).[6] The model outputs a joint posterior distribution of the parameters in the model, as well as statistics for each estimated marginal distribution. In assessing a manipulation's impact on categorization, we report the model estimate and 95% credible interval (CI) for the marginal posterior distribution (where the estimate is the median). An effect is taken to have a meaningful, i.e., credible, influence on categorization when the 95% CI *exclude* zero. An interval encompassing zero would indicate substantial variation in the estimated directionality of the effect, and therefore a non-reliable impact on listeners' categorization.

The model predicted listeners' categorization response ("coat" or "code") as a function of continuum step, pre-target prosodic manipulation, and the interaction of these two fixed effects. The dependent variable was coded with a "code" response mapped to 1. The pre-target prosodic manipulation was contrast-coded with the SHORT condition mapped to –0.5, and the other condition in each sub-experiment mapped to 0.5. Continuum step was treated as a continuous variable and centered at zero. The default prior distribution, an improper uniform distribution over real numbers, was employed in each model. Random effects in each model consisted of by-participant random intercepts and maximal random slopes. Results are shown in Fig. 2. Model summaries are given in Table I.

As shown in Fig. 2, changing pre-target duration only in experiment 1a did not shift reliably listeners' categorization of the target sound. Nor did changing pitch (BT condition) alone in experiment 1b. On the other hand, changing both pre-target duration and pitch in experiment 1c shifted listeners' categorization ($\beta = -0.26$, 95% CI = [−0.49, −0.06]), with the LONG+BT condition showing decreased "code" responses relative to the SHORT condition, showing the predicted proximal rate effect.

This indicates that the expected effect occurs only when pre-target pitch *and* duration are manipulated in tandem, but not when only duration is manipulated, somewhat surprisingly. The lack of the duration effect in experiment 1a presents a departure from the body of literature which documents proximal rate effects, outlined earlier. There might be two reasons for this. One notable difference between the present durational manipulation and those used in Kim and Cho (2013) and Steffman (2019), both of which found an effect of preceding length manipulations on categorization, is the stimulus in the "long" condition: in both of these studies, it was produced at a natural phrasal boundary, with other boundary cues present (i.e., changes in voice quality, intensity, etc., in addition to lengthening). In contrast, in experiment 1a, only the duration was present as a cue. It is thus possible that these additional cues in earlier studies might have helped signal a rate slowdown to listeners. Another difference from the earlier studies is a relatively small durational difference across duration-manipulating conditions in the present study, where the ratio between the durations in SHORT and LONG conditions is 1.6 (125 versus 200 ms). Some of the previously referenced studies included larger

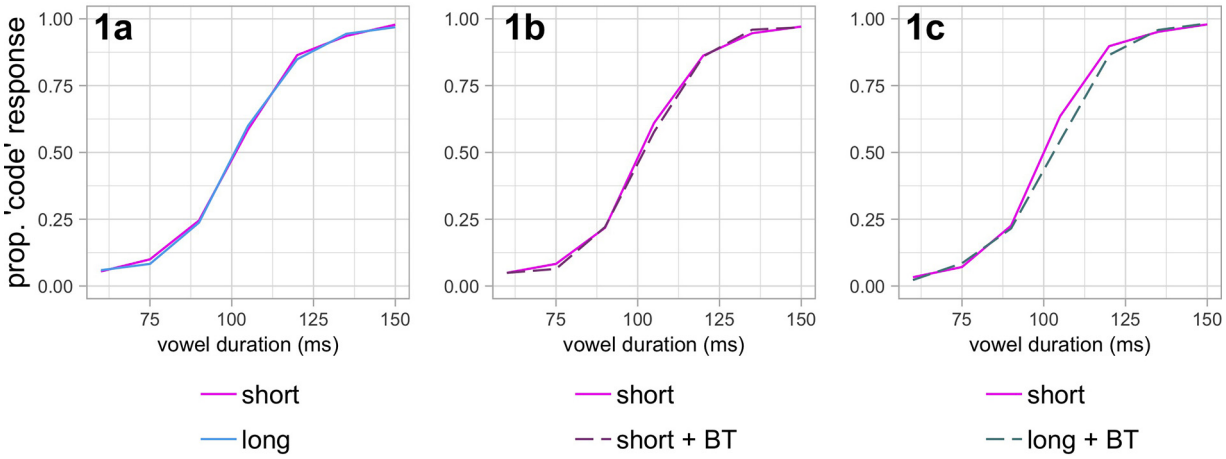

FIG. 2. (Color online) Categorization in experiment 1a (left) and experiment 1b (center), and experiment 1c (right). In each, the *y* axis shows the proportion of "code" responses, along the continuum on the *x* axis. Categorization is split by pre-target prosodic condition, labeled below each plot.

differences across duration-manipulating conditions [e.g., Miller and Volaitis (1989), where the ratio was 2.6, or Summerfield (1981), where the ratio was 2.0]. The present durational differences may thus be insufficient to generate a contrast effect independently, such that an additional cue to a rate slowdown (i.e., pitch) was needed, as seen in experiment 1c. Experiments 2 and 3 will address this point.

Considering the pitch-only variation in experiment 1b, we also do not see a reliable effect, unlike Steffman (2019). In that study, recall a compressed low-rising tune (a L-H% boundary tone in American English intonational phonology) did shift categorization of subsequent temporal contrasts. The present case may seem analogous, but one crucial difference from Steffman (2019) is the actual boundary tone used. The falling tone used here may not be necessarily interpreted as a boundary tone—a low falling contour (over a short vowel) here might be interpreted by listeners as part of a bitonal L + H* pitch accent on the following target word, which would not imply any sort of phrasal boundary on the low toned pre-target vowel. Though this pitch could lead to increased prominence on the target (n.b., L + H* is perceived as more prominent than H* in American English; Cole *et al.*, 2019; Bishop *et al.*, 2020), this is clearly not sufficient to shift categorization of the following target sound. No such interpretation is available for the low-rising boundary tone in Steffman (2019), which was placed on an unaccented syllable and lacked prominence.

However, we can also note that though the effect is not credible, the estimate is negative ($\beta = -0.05$), and 75% of the posterior distribution show a negative sign. Because the posterior in a Bayesian model is a distribution over estimates, we can rephrase this to say that the model estimates a 75% probability that a boundary tone decreases "code" responses to some degree, as calculated with the *p_direction* function in the R package BAYESTESTR (Makowski *et al.*, 2019): at best very weak evidence for an effect of pitch [see, e.g., Makowski *et al.* (2019) for discussion of interpreting posterior distributions in Bayesian modeling]. We note too that the estimate is also quite small, compared to $\beta = -0.26$ in experiment 1c. The directionality of the effect is consistent with the idea that pitch could serve as an independent cue to a rate slowdown [unlike Steffman (2019)] that is enhanced when coupled with longer durations (as in experiment 1c). Experiment 3 will test for this possibility further, as will be described in Sec. IV.

The results of experiment 1 overall suggests that perception of localized rate changes incorporates tonal boundary cues, and additionally, that with our stimuli varying only duration (experiment 1a) or only pitch (experiment 1b) does not produce analogous shifts in categorization of the target as when these cues are combined.

The task in experiment 1 constitutes "implicit" tests of the perception of duration, in the sense that listeners' voicing judgments index their perception of durational differences. Findings so far indicate that these implicit durational processes integrate pitch and duration as cues to rate slowing, as outlined above. However, this task differs from that used in much of the psychoacoustic literature, in which listeners are asked to provide judgements about the duration of acoustic events or intervals between them (e.g., "which is longer?"), using numerical rating or a forced choice

TABLE I. model outputs for experiment 1. "vdur.scaled" stands for (scaled) vowel duration along the continuum. Credible effects are bolded.

		β	Error	L 95% CI	U 95% CI
Exp. 1a	(Intercept)	**0.32**	**0.10**	**0.12**	**0.53**
	pre-target pros	−0.03	0.09	−0.21	0.15
	vdur.scaled	**3.09**	**0.20**	**2.70**	**3.50**
	pros:vdur.scaled	0.08	0.14	−0.19	0.36
Exp. 1 b	**(Intercept)**	**0.41**	**0.12**	**0.17**	**0.64**
	pre-target pros	−0.05	0.07	−0.20	0.09
	vdur.scaled	**3.35**	**0.22**	**2.92**	**3.79**
	pros:vdur.scaled	−0.05	0.15	−0.34	0.25
Exp. 1c	**(Intercept)**	**0.42**	**0.12**	**0.21**	**0.64**
	pre-target pros	**−0.26**	**0.11**	**−0.49**	**−0.06**
	vdur.scaled	**3.76**	**0.30**	**3.19**	**4.37**
	pros:vdur.scaled	−0.14	0.20	−0.56	0.23

https://doi.org/10.1121/10.0007222

comparison (Brigner, 1988; Yu, 2010; Jones and McAuley, 2005). Judgements of duration of this sort can be characterized as "explicit." One question which is unanswered in regards to our present study is accordingly whether the implicit results obtained thus far translate into explicit judgments of duration—that is, whether listeners integrate pitch and durational cues when asked to judge how long the manipulated pre-target word "say" sounds to them. This poses an interesting question because previous studies have found dissociations between these judgment types in rate-dependent perception (Reinisch, 2016) and other domains of cognition more generally (Vorberg *et al.*, 2003). By seeing if the effect observed in experiment 1c replicates in an explicit task, we can glean insight into how listeners are using the tonal and durational information in our stimuli. Experiment 2 addressed this question.

Before reporting experiment 2, a review of Reinisch (2016) is in order, as this study presents several relevant parallels and helps frame the question of differences between implicit and explicit perceptual judgements. Reinisch found a clear dissociation between implicit and explicit judgement tasks when testing overall speech rate differences (the rate of an entire precursor sentence preceding a target sound). In one experiment, listeners categorized a German vowel length contrast preceded by rate that was either fast or slow, and additionally either had, or lacked, fast speech processes (e.g., segmental reductions and deletions). The presence of fast speech processes, independent of, and in addition to, actual rate, shifted listeners' perception of the vowel length contrast, suggesting fast speech processes led to a percept of increased speech rate. However, in a task where listeners were asked to explicitly compare the stimuli by choosing which one sounded faster in a forced choice comparison, the effect of fast speech processes disappeared. Only the actual rate of the sentences affected listeners' explicit judgements. Reinisch notes that this lack of an effect could originate from listeners' taking the speed of segmental articulations and transitions into account in their explicit judgments; when fast speech processes like deletions and reductions are present, they cause a perceived *decrease* in rate as transition speed reduces with deletions. This would compete with the influence of fast speech processes as signaling an *increase* in rate, negating its influence. By this account, the same competition must not take place in implicit judgments, where fast speech processes did matter. Alternatively, implicit and explicit judgments could tap into fundamentally different types of processes: in the implicit task the goal of the participant is to identify the target word, while in the explicit task their goal is to provide a more metalinguistic judgment about some property of the stimulus. Comparison of implicit and explicit metrics of duration perception for localized changes (as in the present study) has, to our knowledge, not been carried out. Accordingly, by using an explicit task we can test if the asymmetry observed in Reinisch (2016) extends to local/proximal rate effects, and test Reinisch's postulation that explicit rate judgments may incorporate some calculation of segments per unit of time, cancelling out competing influences in rate perception. Our stimuli, which only vary in pitch, are matched in terms of segments per unit of time, such that a dissociation observed here would implicate a broader dissociation between implicit and explicit tasks as it pertains to the perception of *localized* durational changes.

III. EXPERIMENT 2

A. Materials

Experiment 2 used the same stimuli as experiment 1 with all conditions (SHORT, LONG, SHORT+BT, LONG+BT), but employed just the endpoints of the target vowel duration continuum. Only the endpoints were used because we wanted listeners to focus on the pre-target syllable, and so did not want to present ambiguous target words. There were accordingly eight unique stimuli [four pre-target prosodic conditions, and two continuum values (the shortest and the longest)].

B. Participants and Procedure

36 new participants were recruited from the same population as experiment 1. The procedure was a simple rating task during which listeners provided numerical ratings to stimuli, using a Likert-style scale labeled with the values 1–5. Listeners were instructed that they would hear a speaker say the sentence "I'll say code now" and "I'll say coat now" and that their task was to focus on the word "say" and rate how long it sounded to them, where 5 indicated the long end point of the scale [e.g., Yu (2010) and Yu *et al.* (2014)]. Participants were told to use the scale as they see fit, and were given eight practice trials that presented each of the unique stimuli in a random order, to give them a sense for the range of variation they could expect during the test trials. Test trials consisted of 12 randomized presentations of each unique stimulus (96 trials total).

C. Results and discussion

The statistical assessment of ratings was carried out by a Bayesian mixed effects ordinal regression on listener ratings [see, e.g., Bürkner and Vuorre (2019)]. The model predicted ratings as a function of contrast-coded pitch and duration (BOUNDARY TONE mapped to 0.5, no boundary tone mapped to −0.5; LONG mapped to 0.5, SHORT mapped to −0.5). Random effects in the model were specified with by-subject and by-word ("coat" or "code") intercepts and fully specified random slopes for subjects. For ease of presenting the results, we plot mean ratings and 95% CI in Fig. 3.

As shown in Fig. 3 and Table II, LONG/SHORT conditions varied in listeners' explicit rating of pre-target duration ($\beta = 1.06$, CI = [0.80,1.31]), while changes in pre-target pitch only did not. That is, listeners rated the word "say" as sounding longer when it was physically longer in duration, as compared to when it was physically shorter. The presence or absence of a boundary tone did not generate any credible

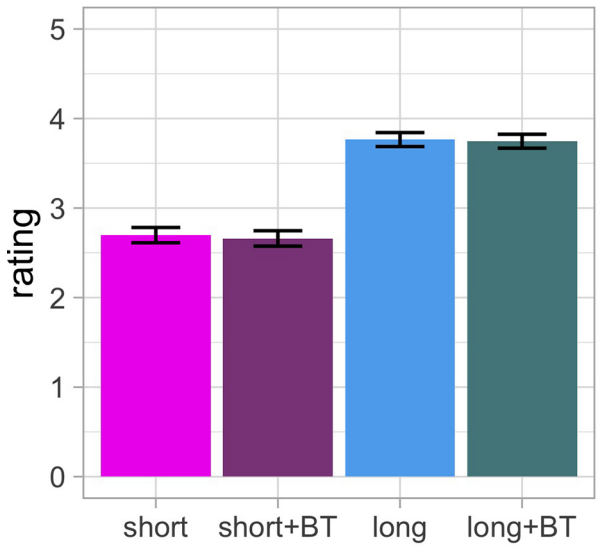

FIG. 3. (Color online) Listeners' mean rating responses (5 = longest) in experiment 2, split by condition, with error bars showing 95% CI.

adjustment in ratings, either as a main effect or an interaction with duration.

We can compare the current findings to experiment 1 in several ways. In experiment 2 the LONG condition, being physically longer in duration, was rated reliably as longer than the SHORT condition. However, in experiment 1a, where listeners were presented with the same LONG and SHORT conditions, no difference across conditions was observed; that is, though the difference in duration is clearly perceptible as indexed by ratings, it did not produce a shift in the categorization of the target sound. In comparison to this, in experiment 1c there was a clear effect of pre-target duration and pitch which shifted categorization of the target sound between the LONG+BT and SHORT conditions, and experiment 2 showed a clear difference in ratings between these two conditions. However, the LONG+BT and LONG conditions showed fairly analogous ratings in experiment 2; that is, the presence of only a boundary tone did not lead to an increase in perceived duration. This lack of an effect of pitch also translates to the two SHORT conditions in experiment 2, aligning with the absence of an effect of pitch in experiment 1b.

These dissociations between implicit and explicit perception of duration are notable for several reasons. First, a clear difference in ratings between the LONG and SHORT conditions does not translate to perceptual adjustments in an implicit task. At the same time, though no difference between the LONG and LONG+BT conditions is found in an explicit task, the LONG+BT condition showed a clear difference from the SHORT condition in an implicit task, though we can only infer that this asymmetry is due to the task, as we did not test the difference between LONG and LONG+BT conditions directly in experiment 1 (this point is addressed in experiment 3). These results are comparable to those found by Reinisch (2016): in that study, listeners' explicit judgments only shifted on the basis of actual differences in rate (here, physical duration), but orthogonal variation of fast speech processes such as segmental reduction or deletion did not impact explicit judgements. Nevertheless, fast speech processes in Reinisch shifted categorization in an implicit task. Unlike Reinisch's manipulations, however, ours did not alter the density of segments per unit of time or articulatory transition speed, and accordingly this does not offer a possible explanation for the absence of an effect of pitch on ratings. Instead, this result suggests that the explicit and implicit tasks presented here are indeed tapping into different processing strategies. Further, when listeners in our experiment were tasked with judging the actual duration of the pre-target word, without the goal of categorizing the target, pitch differences in our stimuli were not sufficient to influence perceived duration, being smaller than those used in previous experiments (Brigner, 1988; Yu, 2010; Yu *et al.*, 2014). If we had only run experiment 2, we might have concluded that pitch is inconsequential and does not influences perception of duration (and by extension durational cues) in this context. However, in comparing experiment 1a and experiment 1c we have some evidence that pitch, as a tonal cue to a prosodic boundary, is consequential in rate-dependent perception. The dissociation between explicit and implicit measures in this regard further suggests that future work investigating the perception of duration and rate-dependent processing might benefit from using both tasks in tandem.

Experiment 1 presents an imperfect test of our question, given that we only examined within-experiment pairwise differences between conditions, which do not allow for a full test of how cues are used in combination (i.e., as would be possible in a fully crossed 2×2 design of duration and pitch). The targeted pairwise comparison across experiment 1a and 1c does still show that pitch plays a role. However, we can only infer this in comparing these two experiments with one another, and further cannot be sure of the relative importance of duration and pitch in the LONG+BT condition in experiment 1c. Experiment 1 thus motivates several changes in our design adopted in experiment 3 (described below) and further provides a useful comparison against experiment 2 as outlined above. One particular point of interest is that durational differences which are insufficient to produce an implicit effect without the presence of a boundary tone (experiment 1a) produce a robust effect when tested in an explicit task (experiment 2). As noted above, the absence of durational contrast effects in experiment 1a could have arisen from a too-small difference across duration conditions. In enlarging the duration difference across conditions, we might obtain a contrast effect in an implicit task. If obtained, we can then ask how this effect interacts with

TABLE II. Fixed effects in experiment 2. Threshold values are excluded from the summary.

	β	Error	L 95% CI	U 95% CI
pitch	−0.03	0.04	−0.11	0.04
duration	**1.06**	**0.13**	**0.80**	**1.31**
pitch:duration	0.02	0.07	−0.13	0.17

https://doi.org/10.1121/10.0007222

pitch, testing if the influence of duration and pitch are additive, and if the influences seen in experiment 1 were evident when listeners hear stimuli varying both in duration and pitch, in a 2 × 2 design. Experiment 3 tested these questions.

IV. EXPERIMENT 3

To address the questions raised above, experiment 3 enlarged the difference in duration between LONG and SHORT conditions, to test if this elicited the predicted contrast effect, and if so, how this effect combines or interacts with the influence of pitch.

A. Materials

Using the same procedure as the original stimulus creation, we modified the LONG and LONG+BT conditions to contain a longer, 300 ms, *pre*-target vowel (the original long conditions in experiments 1 and 2 were 200 ms). These modified LONG conditions in experiment 3 will be referred to by the same names as in previous experiments. We re-used the same SHORT condition stimuli from previous experiments (recall the pre-target vowel in the SHORT conditions was 125 ms, so the ratio across duration conditions is now 2.4). Retaining the same SHORT conditions allows us to attempt to replicate the lack of a credible effect between SHORT and SHORT+BT conditions in experiment 1b. There were a total of 28 unique stimuli (4 conditions, 7 continuum steps).

B. Predictions

Given the enhancement of the proximal durational difference, we predicted we should see an effect of duration, both with and without a boundary tone (comparing SHORT/LONG conditions, and SHORT+BT/LONG+BT conditions). Second, given that we are using the same SHORT and SHORT+BT stimuli as in experiment 1b, we expected to replicate experiment 1b, which would support the idea that pitch alone is not a strong cue when duration is SHORT. We additionally predicted a possible difference between LONG and LONG+BT stimuli, if it is the case that the pitch pattern in the +BT condition contributes to a percept of rate slowing.

C. Participants and procedure

We recruited 80 participants from the same population and based on the same selection criteria to participate remotely in experiment 3.[7] Unlike previous experiments, which were carried out in a lab setting, these participants were instructed to complete the experiment while in a quiet location and wearing headphones. The procedure was identical to experiment 1, however, due to the inclusion of all four duration and pitch conditions we reduced the number of repetitions of each unique stimulus to 10 (from 16 in experiment 1), for a total of 280 randomized trials in the experiment. There were eight randomized practice trials (as compared to four in experiment 1) in which participants heard each continuum end point in each pitch/duration condition. Participants were prompted to take a short self-paced break halfway through the experiment.

D. Results and discussion

Categorization responses were assessed in the same way as in experiment 1, predicting listeners' responses ("coat" mapped to 0, "code" to 1) as a function of pitch and duration (BOUNDARY TONE mapped to 0.5, no boundary tone mapped to –0.5; LONG mapped to 0.5, SHORT mapped to −0.5), target vowel duration, and all interactions. Random effects were specified as in previous experiments, with by-participant intercepts, and slopes for all fixed effects and interactions. Figure 4 plots the results in two ways: at left, responses along the continuum; at right, overall "code" responses, collapsed across the continuum, to better visualize differences across conditions. Table III shows the full model output. Model contrasts were additionally extracted from the pairwise combinations of pitch and duration terms using the package *emmeans* (Lenth *et al.*, 2018). The estimated marginal effects obtained with this method provide the median of the posterior distribution for a given contrast accompanied by 95% highest posterior density credible

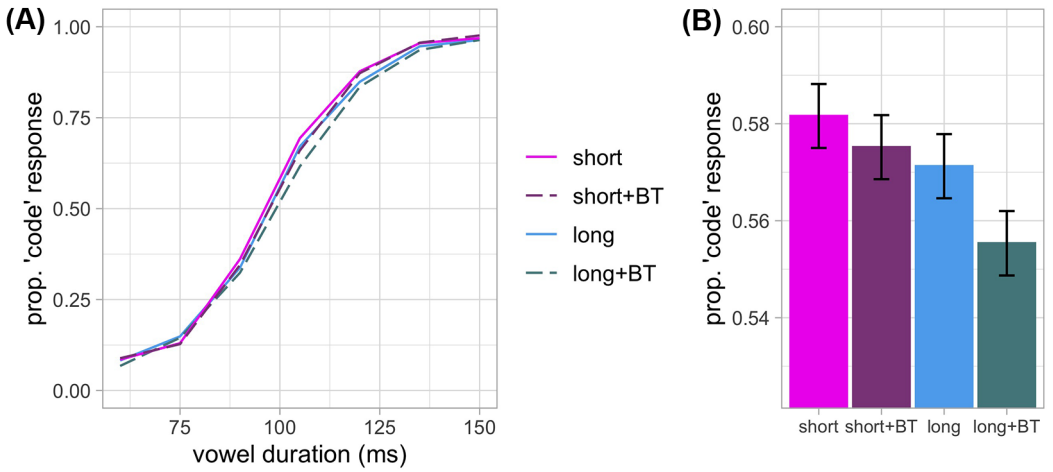

FIG. 4. (Color online) Categorization responses split by condition and continuum step (A) and pooled across continuum steps (B) in experiment 3.

TABLE III. Model output for experiment 3. Credible effects are bolded.

	β	Error	L 95% CI	U 95% CI
(Intercept)	**0.69**	**0.09**	**0.52**	**0.86**
pitch	**−0.10**	**0.05**	**−0.18**	**−0.01**
duration	**−0.19**	**0.05**	**−0.30**	**−0.09**
vdur.scaled	**2.82**	**0.14**	**2.56**	**3.08**
pitch:duration	−0.10	0.09	−0.28	0.08
pitch:vdur.scaled	0.02	0.07	−0.11	0.16
dur:vdur.scaled	**−0.21**	**0.08**	**−0.38**	**−0.05**
pitch:dur:vdur.scaled	−0.02	0.13	−0.26	0.23

intervals, allowing us to assess how a given condition differs from another, shown in Table IV.

In addition to an expected effect of target vowel duration, both the main effect of pitch and the main effect of duration were found to be credible (pitch: $\beta = -0.10$, CI = [−0.18, −0.01]; duration: $\beta = -0.19$, CI = [−0.30, −0.09]). For pitch, the effect shows the presence of a boundary tone (overall) decreases "code" responses. The effect of duration additionally evidenced the predicted contrast effect: a longer pre-target vowel led to decreased "code" responses. With these two main effects we thus have evidence for what could be characterized as an additive influence of pitch and duration, visible in panel B of Fig. 4 which shows a graded decrease in "code" responses, with the SHORT condition showing the most and the LONG+BT condition showing the least. The interaction between pitch and duration showed a skewed posterior distribution but was not credible at 95% CI ($\beta = -0.10$, CI = [−0.28, 0.08]), with 86% of the posterior showing a negative sign. Here, again, in the Bayesian framework we can interpret the posterior distribution as a probability distribution over estimates, such that we can take it as providing weak evidence for an asymmetry across pitch and duration conditions, though this should be interpreted cautiously.

Turning to the pairwise comparisons of interest shown in Table IV, the effect of pre-target duration was credible both with and without a boundary tone present, giving us clear evidence for a durational contrast effect in both cases. The estimated effect is larger however in the boundary tone conditions ($\beta = -0.24$) as compared to the no boundary tone conditions ($\beta = -0.14$), suggesting that when a boundary tone is present, a pre-target slow-down in the LONG+BT condition is more salient, in line with experiment 1, which suggested that a boundary tone and long pre-target vowel, together, cued a rate slowdown. We can also consider the comparison of the LONG+BT and LONG conditions (not tested in experiment 1). Here, too we see a clear effect with the LONG+BT condition showing decreased "code" responses relative to the LONG condition. This further supports the idea that the effects of pitch and duration are additive.

Finally, consider the SHORT+BT condition. Recall that the SHORT and SHORT+BT conditions were the same as in experiment 1b, where we did not see a credible effect of pitch (though we noted the directionality of the effect gave possible weak evidence for pitch as an independent cue to a local rate slow down). The estimate comparing the SHORT and SHORT+BT conditions in experiment 3 is the same as what was found in experiment 1b ($\beta = -0.05$), helping replicate that finding. In the pairwise comparison shown in Table IV, the effect again cannot be deemed credible at 95% CI ($\beta = -0.05$, CI = [−0.17, 0.08], with 76% percent of the distribution having a negative sign). This too is fairly comparable to what was found in experiment 1b ($\beta = -0.05$, CI = [−0.20, 0.09], where 75% of the distribution had a negative sign. However, one piece of evidence in favor of a (small) effect of pitch when duration in SHORT is the fact that the LONG and SHORT+BT conditions are not credibly different from one another, though there is a skewed posterior distribution ($\beta = 0.09$, CI = [−0.04, 0.23]). In other words, the LONG condition is reliably different from the SHORT condition, but not from the SHORT+BT condition. This suggests that the two cues (a boundary tone on a short vowel and an increased duration) lead to a fairly comparable percept of rate slowing and consequent effect on categorization.

To summarize, the results of experiment 3 give us clear evidence for an effect of pre-target duration both with and without a boundary tone, showing the expected contrast effect (though the effect is larger across the two +BT conditions). Second, we have evidence for an additive effect of pitch and duration. This is seen in the main effects of each, the absence of robust evidence for an interaction, and the pairwise difference between the LONG and LONG+BT conditions. We also note that the estimated effect of pitch is substantially larger for the LONG ($\beta = -0.14$) as compared to the SHORT ($\beta = -0.05$) conditions. Following the reasoning outlined above, such an asymmetry in effect size could be attributed to the pairing of the boundary tone and lengthening, where +BT pitch is reliably interpretable as a boundary cue when co-occurring with lengthening and this interpretation is less available in the SHORT condition, resulting in a smaller estimated effect when duration is shorter. In building on the results of experiment 1, these results thus indicate that, even with durational contrast effects present, pitch serves as an additional cue to local rate changes. We have the clearest evidence for pitch as cue to rate slowing when duration is LONG (experiment 1c, and comparisons with the LONG+BT condition in experiment 3).

TABLE IV. Pairwise comparison across conditions in Experiment 3. Credible effects are bolded.

Conditions compared	Estimate	L -95% CI	U -95% CI
LONG+BT - SHORT+BT	**−0.24**	**−0.38**	**−0.10**
LONG+BT - LONG	**−0.14**	**−0.27**	**−0.02**
LONG+BT - SHORT	**−0.28**	**2.56**	**3.08**
SHORT+BT - LONG	0.09	−0.04	0.23
SHORT+BT - SHORT	−0.05	−0.17	0.08
LONG - SHORT	**−0.14**	**−0.28**	**−0.003**

V. GENERAL DISCUSSION

The present experiments explored how listeners incorporate the presence of tonal cues to prosodic structure in

https://doi.org/10.1121/10.0007222

their perception of the temporal patterns in speech, both in terms of their implicit judgement on the categorization of durational contrasts and explicit judgments of stimulus duration. We observed that the presence of a boundary tone, predicted to serve as a tonal cue to local speech rate slowing, engendered, and enhanced (in experiment 3) proximal speech rate normalization effects. In experiment 1, we saw that these effects also did *not* obtain when only duration, or only pitch, varied. However, experiment 3 showed that by increasing the difference in pre-target duration between the SHORT and LONG conditions, we could elicit a durational contrast effect. The effect of pitch, though smaller than the duration effect, was still credible, showing an additive influence with duration, though we remark here again that the effect of pitch was estimated to be numerically larger when duration was LONG.

These results together bear on several broader issues outlined above. First, in building on Steffman (2019), we find that a more general and/or natural pattern of tonal and durational co-occurrence mediates listeners' perception of duration cues in speech, pointing to a relationship between tonal and temporal structure that is exploited by listeners in perception. As outlined in various previous studies, the mediation of perceived rate by some orthogonal factor, in both local and more extended contexts, highlights that rate-dependent perception integrates various sources of information, in line with patterns in the speech signal. The central contribution of the present work is to show that prosodic patterns, and particularly tonal cues, factor into this equation. The present study additionally focused on local speech rate effects. Complementing research which shows that global speech rate effects are driven by more than a syllables-per-unit-of-time calculation of rate (Bosker and Reinisch, 2017; Reinisch, 2016), our findings suggest that local rate perception incorporates multiples cues, above and beyond durational contrast. We are left with a nuanced picture of rate-dependent perception, and our findings point to the importance of considering prosodic organization and cues to prosodic structure in this light. This strengthens recent proposals that rate-dependent perception can be explained both in terms of general auditory mechanisms (contrast, entrainment) as well as linguistic and cognitive factors (Bosker *et al.*, 2017; Bosker and Reinisch, 2017; Reinisch, 2016).

One relevant consideration is cognitive load. It has been shown that cognitive load affects perceived rate: when cognitive load is high, speech sounds faster such that perception of rate-dependent cues shifts (Block *et al.*, 2010; Bosker *et al.*, 2017). This general effect is argued to derive from disruption of temporal sampling on the part of the listener in contexts of higher cognitive effort. Such an explanation has been offered for the observation that non-native speech is perceived as faster, even when rate-matched (Bosker and Reinisch, 2017). Likewise, Reinisch (2016) notes that fast speech processes such as lenitions and reductions might increase cognitive load, leading to more difficulty in recognizing speech and increasing perceived rate. However, our manipulations seem unlikely to introduce variation in cognitive load along these lines: variation in pitch and duration in our stimuli seems unlikely to be harder or easier for listeners to process, and all conditions were created via resynthesis such that slight variations in naturalness in the stimuli should not play a role. Experiment 1 also showed that only changing pitch or duration (if the durational contrast is not large enough) does not reliably shift categorization, something that would be expected if these manipulations somehow influenced ease of processing for listeners.

The present findings raise various questions that will benefit from future research. Various recent studies have pointed to a role for phrasal prosody in speech perception and word recognition [e.g., Kim *et al.* (2018), Mitterer *et al.* (2019), Steffman and Katsuda (2020), and Steffman (2021b)]. In these studies, the authors suggest that listeners make reference to a parsed out prosodic structure which may be integrated in processing as a mediating factor in lexical competition [see also Cho *et al.* (2007)]. The influence of prosodic patterns documented in the present study differs from these studies in that it is more indirect: prosodic structure introduces covariance between acoustic properties (here pitch and duration), which listeners make use of for judging temporal structure in rate-dependent perception [cf. Steffman and Jun (2019)]. Prosodic structure therefore plays an important role, though only in the way it organizes temporal patterns and co-occurrent cues in speech. In this view, perception of local rate changes might incorporate any and all cues that reliably signal prosodic boundaries. A further test of this would be to explore how other cues to a phrasal boundary, such as changes in voice quality and formant structure [e.g., Georgeton *et al.* (2016)] influence rate-dependent perception. It will be useful in this regard to further extend experiment 3 to test if additional cues combine to generate additive shifts in a categorization, and additionally to test if they generate differences in perceived duration as indexed in an explicit task like experiment 2.

Another claim forwarded here is that these influences in rate-dependent perception are derived from patterns in the intonational phonology of American English. The broader implication is that temporal patterns in language, specifically those related to phrasal prosody, should generally shape rate-dependent perception in a language-specific manner. Low or falling pitch and phrase-final lengthening, the two cues tested here, are cross-linguistically common (Cho 2015, 2016; Jun 2005), and accordingly do not present a good test for language-specificity. A further test of this claim would be crosslinguistic extension, testing how temporal patterns more particular to a given language's prosodic system factor in to rate-dependent perception. The present results and those of Reinisch (2016) would suggest that physical/veridical duration and rate play a larger role in explicit judgements, and as such we might expect these to pattern similarly across languages, while implicit judgments might integrate language-specific cues, making language experience more relevant. Seeing if this is indeed the case will help us better understand the mechanisms at play in

both explicit and implicit tasks that test listeners' perception of rate and stimulus duration. Extending the present findings in this way will thus help us better understand how listeners make use of temporal patterns in speech and the ways in which language experience factors into temporal processing.

ACKNOWLEDGMENTS

Many thanks are due to Adam Royer for recording speech for the stimuli, to Danielle Frederickson, Qingxia Guo, and Yang Wang for help with data collection, and to members of the UCLA Phonetics lab for helpful discussion and feedback.

[1]More generally, the relative importance of proximal and distal speech rate contexts is an active area of research [e.g., Bosker (2017), Heffner *et al.* (2013), and Heffner *et al.* (2017)], though we forgo more in-depth discussion here.

[2]The low-rising (L-H%) f0 pattern realized on a short temporal interval retains its boundary tone status in American English, given that a single unaccented syllable with two tonal targets is possible only when the two tones represent a sequence of ip- and IP-final boundary tones (Beckman and Pierrehumbert, 1986). In the long condition, both pitch contours constitute possible boundary tones in American English: a low-rising (L-H%) and a high plateau (H-L%) tones.

[3]Stimuli files can be found in the open access repository for this paper at https://osf.io/qcr62/.

[4]This was done to ensure both conditions were created via resynthesis, in case the PSOLA manipulation decreased naturalness to a small degree. The effect of manipulation would be comparable for all stimuli.

[5]In comparison to the SHORT condition, the SHORT+BT condition has pitch which is both lower on average and also more dynamic, such that it is not *a priori* clear how pitch might influence perceived duration in this case

[6]The data set for each experiment and code used to implement the statistical analysis is available at the online repository for this paper at https://osf.io/qcr62/. Non-Bayesian equivalents of each model reported here were also run and the code for their implementation is in the repository as well. As would be expected, the frequentist implementation of each model finds the same results as the Bayesian version in terms of which effects were and were not credible/significant.

[7]This was due to COVID-19 necessitating remote data collection.

Beckman, M. E., and Pierrehumbert, J. B. (**1986**). "Intonational structure in Japanese and English," Phonology **3**(01), 255–309.
Bishop, J., Kuo, G., and Kim, B. (**2020**). "Phonology, phonetics, and signal-extrinsic factors in the perception of prosodic prominence: Evidence from Rapid Prosody TransCIption," J. Phon. **82**, 100977.
Block, R. A., Hancock, P. A., and Zakay, D. (**2010**). "How cognitive load affects duration judgments: A meta-analytic review," Acta Psychol. **134**, 330–343.
Boersma, P., and Weenik, D. (**2020**). "Praat: Doing phonetics by computer" [computer program], version 6.1.09, http://www.praat.org/ (Last viewed 10/15/2021).
Bosker, H. R. (**2017**). "Accounting for rate-dependent category boundary shifts in speech perception," Atten. Percept. Psychophys. **79**(1), 333–343.
Bosker, H. R., and Reinisch, E. (**2017**). "Foreign languages sound fast: Evidence from implicit rate normalization," Front. Psychology **8**, 1063.
Bosker, H. R., Reinisch, E., and Sjerps, M. J. (**2017**). "Cognitive load makes speech sound fast, but does not modulate acoustic context effects," J. Memory Lang. **94**, 166–176.
Brigner, W. L. (**1988**). "Perceived duration as a function of pitch," Percept. Motor Skills **67**(1), 301–302.
Brysbaert, M., and New, B. (**2009**). "Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English," Behavior Research Methods **41**(4), 977–990.
Bürkner, P.-C. (**2017**). "brms: An R package for Bayesian multilevel models using Stan," J. Stat. Softw. **80**(1), 1–28.
Bürkner, P. C., and Vuorre, M. (**2019**). "Ordinal regression models in psychology: A tutorial," Adv. Methods Practices Psychol. Sci. **2**(1), 77–101.
Chen, M. (**1970**). "Vowel length variation as a function of the voicing of the consonant environment," Phonetica **22**(3), 129–159.
Cho, T. (**2015**). "Language effects on timing at the segmental and suprasegmental levels," in *The Handbook of Speech Production*, edited by M. A. Redford (John Wiley and Sons, New York), pp. 505–529.
Cho, T. (**2016**). "Prosodic Boundary Strengthening in the Phonetics–Prosody Interface," Language Linguistics Compass **10**(3), 120–141.
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," J. Phon. **35**(2), 210–243.
Cole, J., Hualde, J. I., Smith, C. L., Eager, C., Mahrt, T., and Napoleão de Souza, R. (**2019**). "Sound, structure and meaning: The bases of prominence ratings in English, French and Spanish," J. Phon. **75**, 113–147.
Dainora, A. (**2001**). "An empirically based probabilistic model of intonation in American English," Ph.D. dissertation, University of Chicago, Chicago, IL.
Dainora, A. (**2006**). "Modeling intonation in English: A probabilistic approach to phonological competence," in *Laboratory Phonology 8*, edited by L. Goldstein, D. H. Whalen, and C. T. Best (Walter de Gruyter, Berlin).
Diehl, R. L., and Walsh, M. A. (**1989**). "An auditory basis for the stimulus-length effect in the perception of stops and glides," J. Acoust. Soc. Am. **85**(5), 2154–2164.
Dilley, L. C., and McAuley, J. D. (**2008**). "Distal prosodic context affects word segmentation and lexical processing," J. Mem. Lang. **59**(3), 294–311.
Edwards, J., Beckman, M. E., and Fletcher, J. (**1991**). "The articulatory kinematics of final lengthening," J. Acoust. Soc. Am. **89**(1), 369–382.
Epstein, M. A. (**2002**). "Voice quality and prosody in English," Doctoral dissertation, University of California, Los Angeles, CA.
Georgeton, L., Antolík, T. K., and Fougeron, C. (**2016**). "Effect of Domain Initial Strengthening on Vowel Height and Backness Contrasts in French: Acoustic and Ultrasound Data," J. Speech Lang. Hear **59**(6), S1575–S1586.
Green, K. P., Tomiak, G. R., and Kuhl, P. K. (**1997**). "The encoding of rate and talker information during phonetic perception," Perception and Psychophysics **59**(5), 675–692.
Heffner, C., Dilley, L., McAuley, J. D., and Pitt, M. (**2013**). "When cues combine: How distal and proximal acoustic cues are integrated in word segmentation," Lang. Cogn. Process. **28**(9), 1275–1302.
Heffner, C., Newman, R. S., and Idsardi, W. J. (**2017**). "Support for context effects on segmentation and segments depends on the context," Atten. Percept. Psychophys. **79**(3), 964–988.
Jones, M. R., and McAuley, J. D. (**2005**). "Time judgments in global temporal contexts," Percept. Psychophys. **67**(3), 398–417.
Jun, S.-A. (**2005**). "Prosodic typology," in *Prosodic Typology: The Phonology of Intonation and Phrasing*, edited by S.-A. Jun (Oxford University Press, Oxford), pp. 430–458.
Kidd, G. R. (**1989**). "Articulatory-rate context effects in phoneme identification," J. Exp. Psychol.: Human Percept. Perform. **15**(4), 736–748.
Kim, S., and Cho, T. (**2013**). "Prosodic boundary information modulates phonetic categorization," J. Acoust. Soc. Am. **134**(1), EL19–EL25.
Kim, S., Mitterer, H., and Cho, T. (**2018**). "A time course of prosodic modulation in phonological inferencing: The case of Korean post-obstruent tensing," PLoS One **13**(8), e0202912.
Ladd, D. R. (**2008**). *Intonational Phonology* (Cambridge University Press, Cambridge).
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 (Last viewed 10/15/2021).
Liberman, M., and Pierrehumbert, J. (**1984**). "Intonational invariance under changes in pitch range and length," in *Language Sound Structure*, edited by M. Aronoff and R. Oerhle (MIT Press, Cambridge, MA), pp.157–233.
Lisker, L., and Abramson, A. S. (**1970**). "The voicing dimension: Some experiments in comparative phonetics," in *Proceedings of the 6th International Congress of Phonetic Sciences* (Academia Prague, Prague, Czech Republic), Vol. 563, pp. 563–567.
Makowski, D., Ben-Shachar, M. S., and Lüdecke, D. (**2019**). "bayestestR: Describing effects and their uncertainty, existence and significance within the bayesian framework," J. Open Source Softw. **4**(40), 1541.

Makowski, D., Ben-Shachar, M. S., Chen, S. H., and Lüdecke, D. (**2019**). "Indices of effect existence and significance in the Bayesian framework," Front. Psychol. **10**, 2767.
Miller, J. L., Grosjean, F., and Lomanto, C. (**1984**). "Articulation rate and its variability in spontaneous speech: A reanalysis and some implications," Phonetica **41**(4), 215–225.
Miller, J. L., and Volaitis, L. E. (**1989**). "Effect of speaking rate on the perceptual structure of a phonetic category," Percept. Psychophys. **46**(6), 505–512.
Mitterer, H., Cho, T., and Kim, S. (**2016**). "How does prosody influence speech categorization?," J. Phon. **54**, 68–79.
Mitterer, H., Kim, S., and Cho, T. (**2019**). "The glottal stop between segmental and suprasegmental processing: The case of Maltese," J. Mem. Lang. **108**, 104034.
Moreton, E. (**2004**). "Realization of the English postvocalic [voice] contrast in F1 and F2," J. Phon. **32**, 1–33.
Morrill, T., Dilley, L., McAuley, J. D., and Pitt, M. (**2014**). "Distal rhythm influences whether or not listeners hear a word in continuous speech: Support for a perceptual grouping hypothesis," Cognition **131**(1), 69–74.
Moulines, E., and Charpentier, F. (**1990**). "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," Speech Commun. **9**(5–6), 453–467.
Pellegrino, F., Coupé, C., and Marsico, E. (**2011**). "Across-language perspective on speech information rate," Language **87**(3), 539–558.
Pierrehumbert, J. (**1980**). "The phonology and phonetics of English Intonation," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA.
Pitt, M. A., Szostak, C., and Dilley, L. C. (**2016**). "Rate dependent speech processing can be speech specific: Evidence from the perceptual disappearance of words under changes in context speech rate," Atten. Percept. Psychophys. **78**(1), 334–345.
Prince, J. B. (**2011**). "The integration of stimulus dimensions in the perception of music," Quart. J. Exp. Psychol. **64**(11), 2125–2152.
Quené, H. (**2008**). "Multilevel modeling of between-speaker and within-speaker variation in spontaneous speech tempo," J. Acoust. Soc. Am. **123**(2), 1104–1113.
Quené, H. (**2013**). "Longitudinal trends in speech tempo: The case of Queen Beatrix," J. Acoust. Soc. Am. **133**(6), EL452–EL457.
Raphael, L. J. (**1972**). "Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English," J. Acoust. Soc. Am. **51**(4B), 1296–1303.
Reinisch, E., and Sjerps, M. J. (**2013**). "The uptake of spectral and temporal cues in vowel perception is rapidly influenced by context," J. Phon. **41**(2), 101–116.
Reinisch, E. (**2016**). "Natural fast speech is perceived as faster than linearly time-compressed speech," Atten. Percept. Psychophys. **78**(4), 1203–1217.
Redi, L., and Shattuck-Hufnagel, S. (**2001**). "Variation in the realization of glottalization in normal speakers," J. Phon. **29**(4), 407–429.
Šimko, J., Aalto, D., Lippus, P., Włodarczak, M., and., and Vainio, M. (**2016**). "Pitch, perceived duration and auditory biases: Comparison among languages," in *18th International Congress of Phonetic Sciences*, Glasgow, Scotland.
Steffman, J. (**2021a**). "Rhythmic and speech rate effects in the perception of durational cues," Atten. Percept. Psychophys. **83**, 3162–3182.
Steffman, J. (**2021b**). "Prosodic prominence effects in the processing of spectral cues," Language, Cognition Neurosci. **36**(5), 586–611.
Steffman, J. (**2019**). "Intonational structure mediates speech rate normalization in the perception of segmental categories," J. Phon. **74**, 114–129.
Steffman, J., and Jun, S.-A. (**2019**). "Perceptual integration of pitch and duration: Prosodic and psychoacoustic influences in speech perception," J. Acoust. Soc. Am. **146**(3), EL251–EL257.
Steffman, J., and Katsuda, H. (**2020**). "Intonational structure influences perception of contrastive vowel length: The case of phrase-final lengthening in Tokyo Japanese," Lang. Speech. 002383092097184.
Summerfield, Q. (**1981**). "Articulatory rate and perceptual constancy in phonetic perception," J. Exp. Psychol.: Human Percept. Perform. **7**(5), 1074–1095.
Summers, W. V. (**1987**). "Effects of stress and final-consonant voicing on vowel production: Articulatory and acoustic analyses," J. Acoust. Soc. Am. **82**, 847–863.
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, Cogn. Neurosci. **30**(5), 529–543.
Turk, A. E., and Sawusch, J. R. (**1996**). "The processing of duration and intensity cues to prominence," J. Acoust. Soc. Am. **99**(6), 3782–3790.
Turk, A. E., and Sawusch, J. R. (**1997**). "The domain of accentual lengthening in American English," J. Phon. **25**(1), 25–41.
Turk, A. E., and Shattuck-Hufnagel, S. (**2007**). "Multiple targets of phrase-final lengthening in American English words," J. Phon. **35**(4), 445–472.
Turk, A. E., and White, L. (**1999**). "Structural influences on accentual lengthening in English," J. Phon. **27**(2), 171–206.
Vorberg, D., Mattler, U., Heinecke, A., Schmidt, T., and Scharzbach, J. (**2003**). "Different time courses for visual perception and action priming," Proc. Natl. Acad. Sci. **100**, 6275–6280.
Wade, T., and Holt, L. L. (**2005**). "Perceptual effects of preceding nonspeech rate on temporal properties of speech categories," Percept. Psychophys. **67**(6), 939–950.
Welch, T. E., Sawusch, J. R., and Dent, M. L. (**2009**). "Effects of syllable final segment duration on the identification of synthetic speech continua by birds and humans," J. Acoust. Soc. Am. **126**(5), 2779–2787.
Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., and Price, P. J. (**1992**). "Segmental durations in the vicinity of prosodic phrase boundaries," J. Acoust. Soc. Am. **91**(3), 1707–1717.
Wolf, C. G. (**1978**). "Voicing cues in English final stops," J. Phon. **6**, 299–309.
Yu, A. (**2010**). "Tonal effects on perceived vowel duration," in *Laboratory Phonology 10*, edited by C. Fougeron, B. Kuehnert, M. Imperio, and N. Vallee (Walter de Gruyter, Berlin).
Yu, A., Lee, H., and Lee, J. (**2014**). "Variability in perceived duration: Pitch dynamics and vowel quality," *Proceedings of the 4th International Symposium on Tonal Aspects of Languages*, pp. 41–44.