Journal of Phonetics 74 (2019) 114–129

Contents lists available at ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.com/locate/Phonetics

Research Article

Intonational structure mediates speech rate normalization in the perception of segmental categories

Jeremy Steffman

Department of Linguistics, University of California, Los Angeles, 3125 Campbell Hall, Los Angeles, CA 90095, USA

ARTICLE INFO

Article history:
Received 27 May 2018
Received in revised form 26 March 2019
Accepted 27 March 2019
Available online 19 April 2019

Keywords:
Prosody
Intonation
Segmental categorization
Speech perception
Speaking rate normalization

ABSTRACT

The question of if and to what extent listeners' perceptual phonetic categories are sensitive to prosodically driven variability has been a topic of interest in the literature. The present study reports on two experiments which address this question in light of recent research. In Experiment 1, listeners categorized a VOT continuum as /p/ or /b/ in a target syllable (/pɑ/ or /bɑ/). The target was placed in a carrier phrase where the duration and f0 of the pre-target syllable were manipulated. Results suggest listeners are sensitive to intonational structure in their computation of speech rate, interpreting a short syllable with low-rising f0 (created from an L-H% boundary tone in English intonational phonology) as an increase in speech rate. This perceived increase in rate shifts the category boundary of the subsequent target VOT. Experiment 2 showed listeners similarly adjusted categorization of a vowel duration continuum, where vowel duration is a cue to a following obstruent's voicing (categorized as "coat" or "code"). Taken together, these results suggest that listeners are sensitive to intonational structure in their perception of segmental contrasts and use the distribution of tonal targets over a given temporal interval in computing speech rate.

1. Introduction

It has been well established that cross-linguistically, the articulatory and acoustic properties of a speech sound are systematically correlated with the position of that sound in the prosodic configuration of an utterance (e.g. Byrd, 2000; Cho, 2002, 2015; Cho & Keating, 2001, 2009; Fougeron, 1998, 2001, Fougeron & Keating, 1997; Georgeton, Antolík, & Fougeron, 2016; Jun, 1993; Onaka, 2003; Wightman, Shattuck-Hufnagel, Ostendorf, & Price, 1992). However, it remains an open question to what extent this sort of prosodically-driven phonetic detail is relevant in listeners' categorization of speech sounds (Kim & Cho, 2013; Mitterer, Cho, & Kim, 2016). The present study addresses this issue in light of recent research.

The effect of prosodic position on phonetic realization has been conceptualized as the phonetic encoding of prosodic structure (e.g. Keating, 2006), where fine-grained phonetic detail encodes a sound's position in the larger prosodic configuration of an utterance. One well-documented pattern is "initial strengthening" where segments initial to a prosodic unit are realized as "stronger" (e.g. Keating, 2006; Keating, Fougeron, Hsu, & Cho, 2003). "Stronger" can refer to a variety of articulatory and acoustic variables including increased articulatory contact, closure duration, and burst energy (e.g. Cho & Keating, 2009; Fougeron, 2001). The degree of strengthening generally maps hierarchically onto phrasal constituents, where for example, linguopalatal contact was observed to be greater initial to an intonational phrase (IP) than an intermediate phrase (ip) than a word (Fougeron & Keating, 1997). Of specific interest in the present study is how initial strengthening is encoded in the voice onset time (VOT) of voiceless-aspirated stops. In English, VOT is robustly longer when the aspirated stop in question is initial to an IP, versus medial to it (e.g. Cho & Keating, 2009; Pierrehumbert & Talkin, 1992).

Recently, two studies (Kim & Cho, 2013; Mitterer et al., 2016) have investigated how listeners might be sensitive to the relationship between the phonetic realization of VOT and the prosodic position of a segment. Listeners have been shown to shift their perceptual criteria for a given phonetic category on the basis of a variety of contextual factors. These include segmental context (e.g. Harrington, Kleber, & Reubold, 2008; Mann, 1980; Mann & Repp, 1981; Mitterer, 2006), formant frequency distributions preceding the target

E-mail address: jsteffman@ucla.edu

https://doi.org/10.1016/j.wocn.2019.03.002
0095-4470/

(e.g. Holt, Lotto, & Kluender, 2000; Ladefoged & Broadbent, 1957; Sjerps & Smiljanić, 2013) and speaking rate (e.g. Miller & Volaitis, 1989; Newman & Sawusch, 1996; Reinisch & Sjerps, 2013; Summerfield, 1981). Given that phonetic categories are sensitive to this sort of contextual variation, their sensitivity to prosodic factors might be expected. In specific terms, given that VOT is longer in IP-initial position (as compared to IP-medial position), it stands to reason that listeners may take prosodic position into account in their categorization of a VOT continuum, requiring longer VOT for a voiceless stop categorization when the target is IP-initial.

Kim and Cho (2013) carried out a 2AFC (two-alternative forced choice) task experiment addressing this question. Listeners categorized a target sound as /p/ or /b/, in the syllable /pɑ/ or /bɑ/. The target had VOT ranging from 0 to 45 ms in 7.5 ms steps. The crucial manipulation in the experiment was whether the target sound was initial or medial in an IP in the carrier phrase (Kim and Cho's pitch accent manipulation is left aside here; manipulating pitch accent placement did not have a significant effect in their experiment). A ToBI-transcribed representation of this manipulation is shown below, where *x* is the target syllable. In (1) below, the target is IP-medial and in (2) the target is IP-initial.

(1)	Let's H*	hear	*x* H*	again L-L%
(2)	Let's H*	hear L-L%	*x* H*	again L-L%

Kim and Cho found that a post-boundary, i.e. IP-initial, target sound (*x* in (2)) required significantly longer VOT to be categorized as /p/ compared to a target in IP-medial position (i.e. there were decreased /p/ responses in condition (2), compared to (1)). The authors looked at the point in the identification function at which listeners responded /p/ 50% of the time, and found that this point was shifted to higher VOT values by approximately 4 ms in the IP-initial condition. The authors interpreted this effect as originating from speaker sensitivity to the IP boundary and initial strengthening of VOT: "[. . .] when the preceding context provided IP boundary cues, listeners took into account the IP-boundary induced (domain-initial) lengthening of VOT, and therefore required a corresponding, longer VOT for an upcoming post-boundary stop" (p. 24). This interpretation implicates listeners' sensitivity to initial strengthening and prosodic structure in categorization.

However, this interpretation has been reanalyzed more recently by Mitterer et al. (2016), who present speech rate normalization as a possible alternative explanation for the same effect. Generally speaking, speech rate normalization refers to the process by which durational cues to segmental contrasts are relativized to global or local modulations in speaking rate. One notable case demonstrated by Miller, Aibel, and Green (1984) is that the /b/-/w/ boundary, cued by varying transition duration into a following vowel, shifts on the basis of rate information. Relevant to the current study, longer VOT is required for a voiceless percept when in the vicinity of a longer preceding segment (e.g. Summerfield, 1981). Speech rate normalization is a viable alternative explanation because the manipulations represented in (1) and (2) above included changes in duration. In (2) above, the target is preceded by a phrase boundary, manifested in part by pre-boundary lengthening. Given that local modulations in rate exert influence on categorization (e.g. Summerfield, 1981), the lengthened precursor in (2) would be expected to shift the category boundary via rate normalization in the same direction observed by Kim and Cho. Studies that have investigated rate effects on the categorization of VOT have found comparable shifting in 50% crossover points on the basis of rate manipulations as well, suggesting that the effect observed by Kim and Cho might be attributable to rate changes alone (for example, Miller and Volaitis found approximately 8 ms shifting; note their durational difference across rate conditions was slightly larger than Kim and Cho's).

Speech rate normalization is typically viewed as "low-level", domain-general auditory processing, which contrasts with the sensitivity to prosodic structure and its phonetic encoding suggested by Kim and Cho. Arguments for domain generality come from the fact that rate normalization can occur across changes in speaker (e.g. Diehl, Souther, & Convis, 1980; Newman & Sawusch, 2009), and that its effects can be replicated with non-speech analogs (e.g. Diehl & Walsh, 1989; Pisoni, Carrell, & Gans, 1983). Mitterer et al. note that this suggests "[. . .] speaking rate information is used by listeners at a relatively early processing stage which precedes adjustments to speaker differences and auditory perceptual grouping" (p 70). However, the extent to which rate dependent speech perception is an *exclusively* domain-general auditory process is an open question. Numerous studies have shown "higher level" factors related to language experience play a role in rate dependent speech perception (e.g. Baese-Berk, Morrill, & Dilley, 2016; Bosker & Reinisch, 2015, 2017; Dilley, Morrill, & Banzina, 2013,), and other research indicates some rate effects are speech specific (Pitt, Szostak, & Dilley, 2016). These sorts of results suggest that Mitterer et al.'s characterization of speech rate normalization might require more nuance (e.g., Bosker, 2017; Pitt et al., 2016; Wade & Holt, 2005), a point that will be discussed further in Section 5.

Because of the potentially confounding influence of speech rate normalization outlined above, Mitterer et al. performed two experiments aimed at testing Kim and Cho's original claim. In one experiment, Mitterer et al. used the same stimuli as Kim and Cho, but flattened the f0 contour preceding the target (i.e., "let's hear") in (1) and (2) above by setting the pitch to be the mean overall f0 of the utterance. Because prosodic structure is crucially cued by both f0 and duration and because listeners have been shown to be sensitive to f0 modulations for the purposes of word segmentation (e.g. Kim, 2004; Kim & Cho, 2009; Ladd & Schepman, 2003; Spinelli, Grimault, Meunier, & Welby, 2010; Warner, Otake, & Arai, 2010) and syntactic/phrasal grouping and disambiguation (e.g. Kjelgaard & Speer, 1999; Lee & Watson, 2011; Price, Ostendorf, Shattuck-Hufnagel, & Fong, 1991; Schafer, 1996; Streeter, 1978), the authors reasoned that the categorization of stimuli may shift less with a flattened f0 if listeners are compensating for prosodic structure. In other words, if the shift in categorization reflects listeners' sensitivity to prosodic structure, removing one cue to prosodic boundary (here, L-L%) may reduce the magnitude of the effect. However, there was no difference in categorization based on whether the stimuli

had flattened, or naturally contoured, f0: in other words, listeners did not modulate their categorization based on whether the stimuli had flattened pitch. This result led the authors to conclude that the length difference in the precursor between (1) and (2) above was the only factor shifting categorization: " [T]he failure of f0's contribution to boundary-induced modulation of phonetic category may not be seen as entirely consistent with the view that speech perception is directly modulated by computation of prosodic structure" (p. 76). However, as noted by the authors, their result does not entirely preclude listener sensitivity to prosodic structure, as duration serves as a consistent cue to boundary (Shattuck-Hufnagel & Turk, 1996; Tyler & Cutler, 2009) and could have cued a boundary for listeners independently from pitch.

In another experiment, the authors demonstrated that a global slowdown in the precursor "let's hear" shifted categorization in the same direction as the more localized slow-down in (2) above (on "hear"), though the size of the shift was smaller in the globally slowed version. The fact that globally slower speech rate shifted categorization in a similar fashion to pre-boundary lengthening is another piece of evidence that rate normalization may be the driving force behind the effect. The larger effect of localized slowing is explainable by the fact that local durations exert greater influence on upcoming categorization (e.g. Summerfield, 1981), and that listeners normalize over a window surrounding the target segment (e.g. Newman & Sawusch, 1996).

Mitterer et al. therefore suggested that their results are compatible with the hypothesis that listeners are only normalizing for speech rate, though they note the question remains unresolved. In concluding, they state "we are still left with two possible accounts" (p. 70), one in which listeners are sensitive to prosodic patterns in categorizing speech segments, and one in which speech rate normalization is the driving force behind the effect.

1.1. Motivation for the current study

The current study aimed to explore whether intonation independently influences listeners' categorization of speech sounds. Mitterer and colleagues demonstrated that an *absence* of boundary-cueing f0 did not induce any significant shift in categorization compared to the L-L% boundary tone. However, this methodology does not test for the possibility that other f0 contours that uniquely cue an IP boundary might shift categorization. Further, in removing f0 information the authors precluded the possibility of testing whether speech rate normalization is affected by tonal contours, a point that will be discussed further below.

The experimental design independently crossed f0 (intonation) and durational cues to an IP boundary in a 2×2 design, using a similar 2AFC task as that implemented by Kim and Cho (2013) and Mitterer et al. (2016). Specifically, the experiment used an intonation-based cue to prosodic boundary that should persist even in the absence of durational cues. The 2×2 manipulations were made to the syllable immediately preceding the target in the carrier phrase "I'll say *x* again", where *x* is the target (described in more detail in Section 2.1 below). The two durational conditions used are named SHORT and LONG, where the LONG condition presented a durational cue to boundary (phrase-final lengthening), and the SHORT condition did not cue a boundary in terms of duration. These two durational conditions were crossed with two intonation conditions.

The selection of intonational variables for use in the experiment was informed by their predicted interpretation within the framework of English intonational phonology (e.g. Beckman & Pierrehumbert, 1986; Ladd, 2008; Pierrehumbert, 1980). In the intonational phonology of Mainstream American English (MAE) an IP-final syllable can have four possible (non-downstepped) f0 contours:[1] a high rise (H-H%, typical rising intonation in yes/no questions), a high flat plateau (H-L%, used in listing items), a low rise (L-H%, typically described as signaling a continuation), and a low fall (L-L%, used in declaratives). The tones with a hyphen diacritic (H-, L-) are the boundary tones of an intermediate phrase (ip) and can occur over multiple syllables between the nuclear pitch accented word and the IP-final syllable. However the tones with % (H%, L%) are IP boundary tones, and occur only on the IP-final syllable. When both ip and IP boundary tones occur on one syllable, a low-rise (L-H%) is likely to signal that the syllable is IP-final. This is because the L-H% is the only boundary tone that changes the direction of pitch movement, from a low target to a high target, within the same syllable. Therefore, when a syllable carrying both ip and IP tones is not lengthened, all the boundary tones *except* for L-H% might possibly be reinterpreted as a non-boundary tone. For example, low falling f0 (L-L%) over a shortened unaccented syllable could be reinterpreted as a low leading tone for the following H^* pitch accent (an $L+H^*$ pitch accent). A high flat f0 (H-L%) on an unaccented syllable could be interpreted as a non-target transition or sag between H^* pitch targets (e.g., between "*Let's*" and "*x*" in (1)). Similarly, super-high f0 (H-H%) could be interpreted as the H leading tone of the following pitch accent, i.e., $H+!H^*$, or a delayed peak of the preceding H^*. In contrast, two tonal targets (L-H%) on one non-prominent syllable is unlikely to be interpreted as a pitch accent, or a transition tone, or a delayed tone. For example, the low target is aligned too early in the syllable to be interpreted as a preceding L target in an $L+H^*$ pitch accent on the following target sound.

To summarize: in English intonational phonology, an unaccented syllable with two tonal targets has to be an IP-final syllable carrying a L-H% boundary tone. Following this logic, a syllable with a low-rise f0 contour is likely to be perceived as an L-H% IP-boundary tone by listeners, even in the absence of durational cues. For this reason, this L-H% f0 contour was one of the two intonation conditions used in the experiment. Listener interpretation of this contour will be further discussed in Section 3 light of the experimental results.

The second intonation condition used a high flat f0 contour. As mentioned earlier, this contour is not expected to be interpreted as cueing a boundary in the SHORT condition. However, in the LONG condition, it should be interpretable as a boundary tone (H-L%). These two intonation conditions are named the LH and FLAT condition, respectively. Importantly, the two intonation conditions present an asymmetry in how they are

[1] Audio examples and explanations of different MAE ToBI-labeled utterances that exemplify the contours discussed here can be found at https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-911-transcribing-prosodic-structure-of-spoken-utterances-with-tobi-january-iap-2006/ (Veilleux, Shattuck-Hufnagel, & Brugos, 2006).

predicted to be interpreted by listeners in the SHORT condition. This is schematized in Table 1.

As shown in Table 1, in the SHORT condition only, the intonational contour should sound like a boundary tone when LH, but not when FLAT (cf. cell (a) and cell (c) in Table1). The pertinent question is then if and how this will impact categorization of the following target sound.

Table 2 lays out the predicted effects of the manipulations used in the experiment for the two competing accounts discussed above (rate normalization versus compensation for initial strengthening). Note, because pre-boundary lengthening is a consistent cue to boundary (e.g. Klatt, 1976; Shattuck-Hufnagel & Turk, 1996; Tyler & Cutler, 2009; Wightman et al., 1992) and based on Mitterer et al.'s findings outlined above, it is assumed that even under a prosodic compensation account (following Kim & Cho, 2013), the LONG condition will provide a duration-based cue to boundary, triggering compensation for initial strengthening. In other words, duration, independent of intonation, is predicted to shift categorization, either by cueing a boundary, or via speech rate normalization. For that reason, the effect of length is ignored in Table 2.

As shown in Table 2, the critical comparison for teasing apart the two explanations is the effect of intonation in the SHORT condition, where the two accounts make different predictions about if and how categorization will shift. Accordingly, this comparison will be the focus of interpreting the results. Note that because both intonational variables are possible boundary tones when LONG, categorization would not be expected to shift on the basis of intonation in the LONG condition by either of the explanations. Further discussion and exploration of listener interpretation of the stimuli will be carried out in Section 4.

In controlling for the effects of duration and intonation in the manner outlined above, this experimental design offers a way to tease apart listener sensitivity to prosodic structure and normalization for speech rate, building on the work done by Kim and Cho (2013) and Mitterer et al. (2016).

Table 1
Conditions used in the experiment, with predicted listener interpretation.

		DURATION	
		SHORT condition	LONG condition
INTONATION	**Low-rising f0 (LH)**	(a) IP boundary with an L-H% boundary tone (even when SHORT)	(b) IP boundary with an L-H% boundary tone
	High-flat f0 (FLAT)	(c) Not interpretable as a boundary tone	(d) IP boundary with an H-L% boundary tone

Table 2
Predictions for the compared conditions from Table 1, split by account.

	Compensation for initial strengthening (Kim & Cho, 2013)	Rate normalization (Mitterer et al., 2016)
(i) Effect of intonation in the SHORT condition	LH cues a boundary, but FLAT does not. Fewer /p/ responses when LH	No effect, duration is the same
(ii) Effect of intonation in the LONG condition	No effect, both duration and intonation cue a boundary	No effect, duration is the same

(3)	I'll H*	say	*pa* H*	again L-L%
(4)	I'll H*	say L-H%	*pa* H*	again L-L%

2. Experiment 1

Following Kim and Cho (2013) and Mitterer et al. (2016), the experiment was a 2AFC task, wherein participants categorized the target as /p/ or /b/. The platform used for presentation of the stimuli was Appsobabble (Tehrani, 2015).

2.1. Materials

The stimuli used in the experiment were created by resynthesizing the speech of a ToBI-trained English speaker. The procedure for creating the stimuli is outlined below.

The speaker was first recorded at 44.1 kHz (32 bit) using SM10A Shure™ microphone and headset in a sound attenuated room in the UCLA Phonetics Lab. Manipulation was carried out with PSOLA resynthesis (Moulines & Charpentier, 1990) in Praat (Boersma & Weenik, 2018). Two ToBI-transcribed utterances that served as the starting point for stimuli creation are represented in (3) and (4) below. The target word was produced as [p^{h}ɑ] during recording, written as '*pa*' below.

The creation of the stimuli proceeded as follows. First, the vowel in "say" was excised from (4) above. The remainder of (4) served as the frame for all stimuli. Because phrasal boundaries affect the duration of segments that are not directly adjacent to them (e.g. Turk & Shattuck-Hufnagel, 2007), having a frame that was phrased originally as two separate IPs could potentially bias responses. To minimize this possibility, the frame was resynthesized so that the duration of each individual segment was the mean duration of that segment in (3) and (4).

The vowel for the SHORT conditions were created as follows. The vowel in "say" from (3) above (with a duration of 145 ms) was excised and inserted into the frame mentioned above. This vowel had high flat f0 as the interpolation between adjacent H* pitch accents. This created condition (c) in Table 1 (FLAT f0 on a SHORT vowel). Next the contour from the excised vowel in (4) (L-H%) was overlaid onto this SHORT vowel and inserted into the frame, creating condition (a) in Table 1 (LH pitch on a SHORT vowel). The LONG conditions were created as follows. The LONG vowel from (4), with naturally produced L-H% and a duration of 255 ms was reinserted into the duration-normalized frame to create condition (b) in Table 1 (LH pitch on a LONG vowel). The high flat contour from the vowel in (3) was overlaid onto this LONG vowel as well and the vowel was reinserted into the frame to create condition (d) in Table 1 (FLAT f0 on a LONG vowel). Fig. 1 shows the pitch tracks and waveforms for the four conditions.

The duration of the silent interval between the target sound and preceding "say" (which is interpreted as closure duration of /p/) was also manipulated to be different in the LONG and SHORT conditions. In the SHORT condition, the silent gap was set to be 80 milliseconds, and in the LONG condition, it was set to be 140

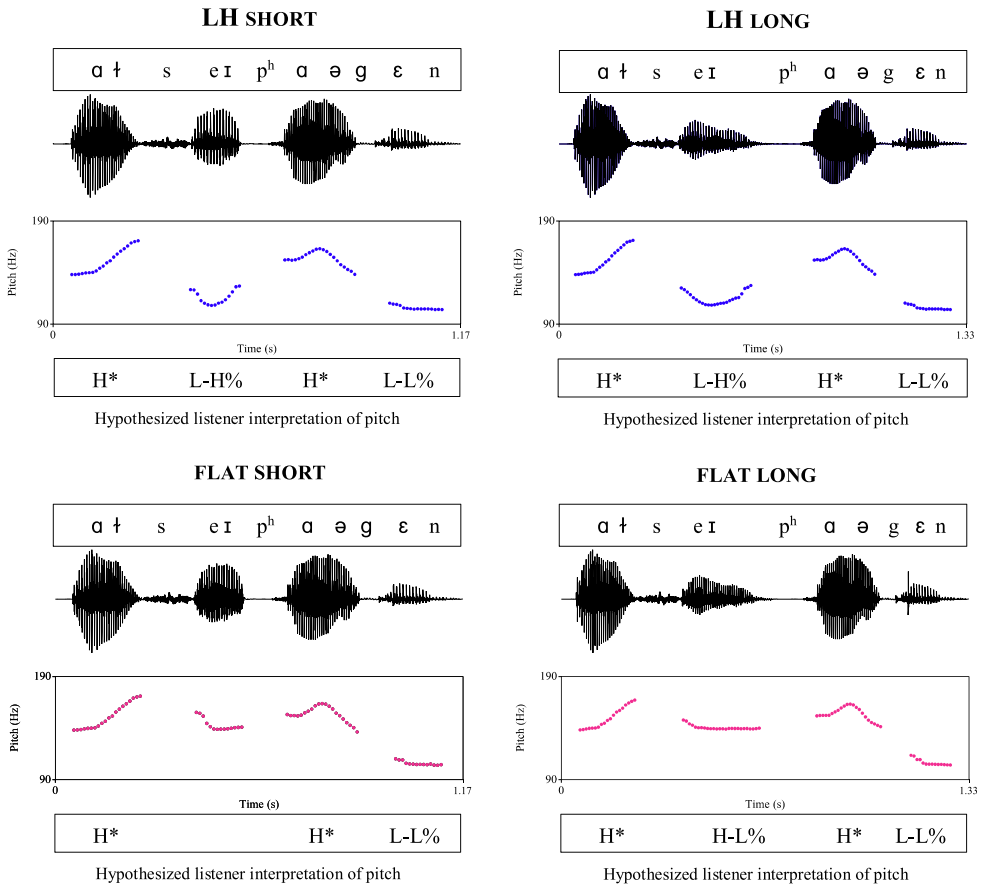

Fig. 1. Waveforms and temporally-aligned pitch tracks of the stimuli sentence "I'll say pa/ba again", for 2 × 2 intonation and length conditions. Below each pitch track a ToBI transcription based on *predicted* listener interpretation is given (see Table 1). A segmental transcription is given above each waveform. The stimuli shown have 45 ms of VOT for the target stop. The pitch range is given on the *y*-axis of each pitch track, and approximate times are given on the *x*-axis. Example sound files are described and linked in Appendix B.

milliseconds. These are similar to the values used by Kim and Cho (2013), shortened slightly to be in line with the natural productions of the speaker who produced (3) and (4) above. This was done so that the difference in the silent interval between the LONG and SHORT conditions was consistent with the level of boundary cued by duration. Longer stop closures occur IP-initially (e.g. Cho & Keating, 2009), and so pairing longer closures with the LONG condition should reinforce an IP-boundary percept for listeners. In the SHORT condition, shorter closure durations are consistent with the absence of a boundary. Because the SHORT condition is where an effect of f0 is predicted, by keeping the silent interval short in this condition it is assured that it will not contribute to a boundary percept, allowing a more direct interpretation of the effect of f0. Importantly, if this difference in closure duration between the SHORT and LONG conditions were to bias responses, it would be expected to bias towards /p/ in the LONG condition, as /p/ has longer closure than /b/ (e.g. Lisker, 1986). The results indicate clearly that this is not the case, showing that the silent interval manipulation is not overriding the effect of length. This is in line with Kim and Cho (2013), who manipulated closure duration orthogonally to other variables and found no significant main effect on categorization.

VOT manipulations were made by resynthesizing the naturally produced duration of the VOT in [pʰ] in (4) (with a duration of approximately 60 ms), shortening the duration with resynthesis to create ten steps on a continuum from 0-45 ms in 5 ms steps. These manipulations resulted in 40 unique stimuli (2 intonation conditions × 2 length conditions × 10 VOT steps).

2.2. Participants

55 self-reported monolingual speakers of American English with normal hearing participated in the study. All participants were students at UCLA and received course credit for participation. Participants were excluded if their mean proportion of /p/ responses fell more than two standard deviations outside the mean proportion of /p/ responses for the group at either endpoint. Four participants were excluded on this basis. One other participant was excluded due to failure to perform the experimental procedure (this participant fell asleep). The results reported here are for the remaining 50 participants (32 identifying as female, 18 identifying as male).

2.3. Procedure

The testing was carried out in a sound-attenuated room in the UCLA phonetics lab. Participants completed testing while

seated in front of a desktop computer. Stimuli were presented binaurally via a Peltor™ 3M™ listen-only headset, with the volume adjusted to a comfortable listening level. Before testing began, participants were told they would listen to a native English speaker saying "I'll say *x* again", and their task was to categorize *x* as beginning with /p/ or /b/.

To familiarize them with the experimental procedure and the stimuli, participants first completed several practice trials in which they heard the endpoints of the continuum and were told that the speaker was saying /bɑ/ (when the 0 ms endpoint was played), and saying /pɑ/ (when the 45 ms endpoint was played; they were not told that these tokens were the endpoints of a continuum). They pressed the appropriate key on the keyboard for each sound. They heard 8 tokens of each endpoint, with two of each of the four conditions from Table 1. Tokens were randomized within this 8 token block. It was also random whether the 0 ms endpoint block or the 45 ms endpoint block was heard first. During testing, participants heard a stimulus and were presented visually with "p" and "b" on the computer screen, one on each side of the screen. Participants indicated their choice via a key press on the computer keyboard, where an 'f' key press indicated the left side choice, and a 'j' keypress indicated a right side choice. The side of the screen on which "p" and "b" appeared was counterbalanced across participants; for 25 "p" was on the left, for 25 "p" was on the right.

Stimuli were organized by the set of 40 unique stimuli, and randomized within this set, meaning that participants heard a randomized mix of the intonation and length conditions. Participants heard 10 such sets total during the experiment, so that each participant categorized a total of 400 stimuli. Participants were given the opportunity to take a short self-paced break halfway through. The inter-trial-interval (time between the key-press response and the beginning of the subsequent stimulus) was 250 ms. The run time for the experiment was approximately 20–25 min.

2.4. Results

In this section the results from Experiment 1 will be outlined with reference to the statistical model used in their evaluation. The results from Experiment 1 were assessed by a linear mixed-effect model with a logistic linking function, using the lme4 package in R (Bates, Maechler, Bolker, & Walker, 2015),[2] to account for the categorical nature of the dependent variable (e.g. Jaeger, 2008). The contrasts in the model were effect-coded (as in Mitterer et al., 2016), for two-level categorical variables (e.g. Bech & Gyrd-Hansen, 2005). Fixed effects specified in the model were VOT (treated as continuous and centered at zero), two levels of intonation (LH and FLAT) and two levels of length (LONG and SHORT), as well as all possible interactions. In contrast coding the categorical fixed effects, FLAT was mapped to 1 and LH was mapped to −1; LONG was mapped to 1 and SHORT was mapped to −1. The random effect structure of the model consisted of by-subject random intercepts, with the maximal number of random slopes that allowed for the model to converge (e.g. Barr, Levy, Scheepers, & Tily, 2013).[3] Data visualization was carried out in RStudio (RStudio Team, 2016). A plot of the categorization functions split by all four conditions is shown in Fig. 2.

Recall the effect of intonation in the SHORT condition is key in probing for listener sensitivity to prosodic structure. Before this particular effect is discussed the other significant effects in the model will be noted.

As would be expected with any such VOT continuum, increasing VOT significantly increased /p/ responses ($B = 2.74$; $z = 18.07$; $p < 0.001$).

Length also showed a significant main effect. As expected, a LONG preceding vowel significantly decreased /p/ responses ($B = -0.27$; $z = -5.44$; $p < 0.001$), shown in Fig. 2. Length also showed a significant interaction with VOT ($B = 0.30$; $z = 6.89$; $p < 0.001$), whereby as VOT increased the effect of length diminished along the continuum. This effect can also be observed in Fig. 2, where categorization by length is separated more at the lower values of the continuum. This indicates that these lower VOT values are more ambiguous to listeners, and thus more susceptible to shifts based on preceding duration.[4]

The intonation manipulations also significantly shifted categorization, where the FLAT condition significantly decreased /p/ responses ($B = -0.06$; $z = -2.41$; $p = 0.016$). This effect is clearly of smaller magnitude than that of length, and is further linked to a significant three way interaction between VOT, intonation, and length ($B = -0.08$; $z = -2.41$; $p = 0.016$). To further investigate the interaction, post-hoc comparison of contrasts with emmeans was used (Lenth, Singmann, Love, Buerkner, & Herve, 2018), where the effect of intonation in each length condition was compared at each VOT value along the continuum (shown in Table 4 below). The test revealed that there was no effect of intonation in the LONG condition at any point along the continuum.

In contrast, in the SHORT condition, intonation had significant effect at the lower end of the continuum. At the five lowest steps on the continuum (0–20 ms VOT), there was a significant effect of intonation, and from 25-45 ms there was no effect. This asymmetry is visible in Fig. 2, where in the SHORT condition only, the lines of the categorization function are consistently separated at the lower end of the continuum. Note that the SHORT LH condition has more /p/ responses as compared to the SHORT FLAT condition. In comparison, no such separation exists along the categorization function in the LONG condition.

Table 3 shows the full output from the statistical model. Table 4 shows the output of the comparison of contrasts with

[2] A full readout of the R code used for all models is given in the appendix.

[3] The model started with all fixed effects and interactions as by-subject random slopes, and was simplified until it converged. Model simplification started by removing correlation parameters for random slopes and then refitting the model (e.g. Bates, Kliegl, Vasishth, & Baayen, 2015; Barr et al., 2013; Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017). Following this, simplification of random slopes started with the highest order random slope. In fitting the model for Experiment 1, correlation parameters for the random slopes and the highest order random slope (VOT:intonation:length, where a colon indicates an interaction) were removed. The model converged with this simplified random effect structure. The remaining by-subject random slopes specified in the model were therefore VOT, intonation, length, VOT:length, VOT:intonation, and length:intonation.

[4] This ambiguity at lower VOT values is likely attributable to the fact that the base for the creation of the VOT continuum was an aspirated [pʰ]. Voice quality and pitch at vowel onset, as well as a lack of prevoicing during closure are all potential cues that rendered these lower values ambiguous to listeners, effectively biasing the continuum towards /p/ responses. A comparable /p/ bias is found in Kim and Cho (2013).

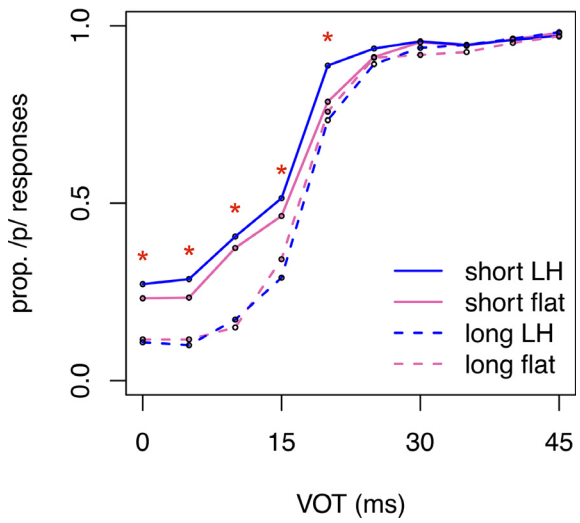

Fig. 2. Experiment 1 categorization, split by all four conditions. The *x* axis shows the VOT values from the continuum, while the *y* axis shows the proportion of /p/ responses at each value. Asterisks above points on the VOT continuum index where a significant difference was found between the SHORT FLAT and the SHORT LH condition, as determined by emmeans comparison of contrasts shown in Table 4. Note no significant difference based on intonation was found at any point along the continuum between LONG conditions.

Table 3
The model output for Experiment 1. Estimates are rounded to two decimal places. Approximate *p* values are shown at right.

	B (SE)	z value	*p* value
(Intercept)	1.49 (0.11)	13.38	<0.001
VOT	2.74 (0.15)	18.07	<0.001
Intonation	−0.06 (0.03)	−2.41	0.016
Length	−0.27 (0.05)	−5.44	<0.001
VOT:intonation	−0.01 (0.04)	−0.29	0.77
VOT:length	0.30 (0.04)	6.89	<0.001
intonation:length	0.04 (0.03)	1.49	0.13
VOT:intonation:length	−0.08 (0.03)	−2.41	0.016

emmeans, showing the asymmetry across length conditions revealed by the three-way interaction.

2.5. Experiment 1 discussion

The results of Experiment 1 highlight how both durational and intonational information modulated categorization of the VOT continuum. The effect of duration, where the LONG condition decreased /p/ responses, is analogous to the effect found by both Kim and Cho (2013) and Mitterer et al. (2016), where a lengthened precursor shifted the categorization to higher VOT values for a voiceless /p/ response. As outlined by Mitterer et al., this effect could simply be normalization for speech rate, or could reflect listeners sensitivity to, and compensation for, initial strengthening. Thus the effect of duration reported here does not serve to tease these potential explanations apart, though it does replicate the effect observed in the studies mentioned above.

The asymmetry in the effect of intonation across length conditions is in line with the predictions shown in Table 1 above. The absence of an effect of intonation in the LONG condition is expected given that both contours are possible boundary tones. An effect of intonation in the SHORT condition is expected because only LH intonation should be interpretable as a boundary tone in this condition (see Table 1).

However, the directionality of the effect is contrary to predictions. Recall that following Kim and Cho (2013), assuming the SHORT LH condition is perceived as containing an intonational cue to boundary by listeners, if listeners expect post-boundary lengthening of VOT (in initial strengthening) they would expect *longer* VOT in the SHORT LH condition relative to the SHORT FLAT condition. This would engender the compensatory effect discussed by Kim and Cho, where categorization in the SHORT LH condition would be shifted to *longer* VOT values (i.e. longer VOT required for a /p/ response), *rightwards* visually in the categorization function. This predicted effect

Table 4
Output from comparison of contrasts with emmeans, investigating the three-way interaction. The effect of intonation in the SHORT condition and LONG condition, at each VOT value is compared. In each VOT grouping, the SHORT condition is the first row, and the LONG condition is the second row. Significant *p* values are bolded.

VOT	Conditions compared	Estimate (SE)	z ratio	*p* value
0 ms	SHORT LH – SHORT FLAT	−0.41 (0.13)	−2.98	**0.015**
	LONG LH – LONG FLAT	0.24 (0.16)	1.47	0.45
5 ms	SHORT LH – SHORT FLAT	−0.36 (0.11)	−3.36	**0.004**
	LONG LH – LONG FLAT	0.17 (0.13)	1.38	0.51
10 ms	SHORT LH – SHORT FLAT	−0.32 (0.08)	−3.79	**<0.001**
	LONG LH – LONG FLAT	0.11 (0.09)	1.17	0.65
15 ms	SHORT LH – SHORT FLAT	−0.27 (0.07)	−3.94	**<0.001**
	LONG LH – LONG FLAT	0.05 (0.08)	0.69	0.90
20 ms	SHORT LH – SHORT FLAT	−0.23 (0.07)	−3.25	**0.006**
	LONG LH – LONG FLAT	−0.01 (0.07)	−0.14	0.99
25 ms	SHORT LH – SHORT FLAT	−0.18 (0.09)	−2.14	0.14
	LONG LH – LONG FLAT	−0.07 (0.08)	−0.84	0.83
30 ms	SHORT LH – SHORT FLAT	−0.14 (0.11)	−1.27	0.58
	LONG LH – LONG FLAT	−0.13 (0.11)	−1.20	0.63
35 ms	SHORT LH – SHORT FLAT	−0.10 (0.14)	−0.69	0.89
	LONG LH – LONG FLAT	−0.19 (0.14)	−1.36	0.53
40 ms	SHORT LH – SHORT FLAT	−0.05 (0.17)	−0.31	0.99
	LONG LH – LONG FLAT	−0.25 (0.18)	−1.44	0.47
45 ms	SHORT LH – SHORT FLAT	−0.01 (0.20)	−0.04	0.99
	LONG LH – LONG FLAT	−0.31 (0.21)	−1.48	0.45

would therefore translate to *decreased* /p/ responses overall in the SHORT LH, as compared to SHORT FLAT condition (see Table 2).

In fact, the opposite directionality of the effect is observed. As shown in Fig. 2, in the SHORT condition only, LH intonation overall *increased* /p/ responses, with a significant effect at lower, more ambiguous, VOT values. Visually, categorization in the SHORT LH condition is shifted *leftwards*, to *shorter* VOT values on the continuum.

These unexpected results (increased, not decreased /p/ responses) suggest a novel interpretation: that listeners are using intonational information to compute speech rate. The logic of this interpretation is as follows. Given that L-H% naturally co-occurs with phrase-final lengthening, when the contour was overlaid on a non-lengthened segment it appears to be interpreted as faster speaking rate, shifting categorization to a lower VOT value for a /p/ response in the following target sound. In other words, because listeners are sensitive to the fact that L-H% co-occurs with phrase final lengthening, when it was compressed onto a non-lengthened segment, it gave the impression of increased speech rate. This perceived increase in speech rate in the SHORT LH condition then altered subsequent categorization of VOT such that listeners shifted categorization to *shorter* VOT values for a /p/ response, as would be expected (e.g. Summerfield, 1981). This is in contrast to the SHORT FLAT contour, which is argued to be a typical transition between adjacent pitch accents, as discussed above. The SHORT FLAT condition would therefore *not* generate a perceived increase in rate. This interpretation also predicts the absence of an effect in the LONG condition, as was observed in Experiment 1. This is for the simple reason that both tonal contours are plausible boundary tones (see Table 1), both occurring with lengthening. Because of this, neither would be expected to give the impression of a change in speech rate when co-occurring with a LONG vowel.

In other words, the perception of an intonational cue to boundary could possibly have two different effects on categorization: one in which post-boundary compensation for initial strengthening occurs (Kim & Cho, 2013; predicted in Table 2), or one in which the tonal cues to boundary generate a perceived increase in speech rate, which modulates post-boundary categorization. The results of Experiment 1 are consistent with this latter effect, which is further explored in Experiment 2.

To summarize, if this interpretation is correct, the results of the first experiment indicate that intonation is relevant to listeners in computing speech rate, which in turn modulates categorization of speech segments. This account therefore still holds that prosodic/intonational structure is relevant in listeners' perception of segmental categories, as argued by Kim and Cho, but suggests that the intonational cues used in this study modulate listeners' perception of rate, as opposed to their expectations about post-boundary lengthening of VOT.

The suggestion that intonational structure influences how speakers perceive speech rate has in fact been made before. Rietveld and Gussenhoven (1987) showed that listeners' rate judgments are influenced by the intonational structure of utterances in Dutch, though they tested unrelated intonational variables and used explicit rate judgment tasks. Explicit judgments of rate differ from the present study's use of segmental categorization which tests implicit rate normalization. Research has also shown implicit speech rate normalization and explicit judgments do not necessarily align (Reinisch, 2016). Accordingly, Rietveld and Gussenhoven's results are broadly compatible with the central argument made here, though they do not offer direct support for the present findings. In order to strengthen the claim that intonational structure mediates the perception of speech rate, a second experiment was carried out.

3. Experiment 2

If listeners use intonation to compute speech rate and adjust categorization of speech sounds, then the effect observed in Experiment 1 would be expected to generalize to other durational contrasts. To test this, a second experiment was carried out with a vowel duration continuum, where listeners used the duration of the vowel to categorize a following stop as voiced or voiceless.

In English, vowels are significantly longer preceding voiced (as opposed to voiceless) obstruents (e.g. Chen, 1970; Walsh & Parker, 1981), and in perception, listeners use preceding vowel duration as a cue to obstruent voicing (e.g. Raphael, 1972). In Experiment 2, participants categorized the target as one of two lexical items, "coat" or "code", where the endpoint of the continuum with the shortest vowel duration should be categorized as "coat" and the endpoint with the longest duration should be categorized as "code". These particular words were chosen because they have relatively similar lexical frequencies ("code" $\text{Log10}_{\text{WF}} = 3.43$; "coat" $\text{Log10}_{\text{WF}} = 3.33$), as obtained from the SUBTLEXUS corpus (Brysbaert & New, 2009). Controlling for frequency in this way minimizes a potential frequency bias in categorizing the continuum.

The crucial manipulations in Experiment 2 were identical to those used in Experiment 1, with the same 2×2 conditions shown in Table 1. Based on the results of Experiment 1, the following predictions were made. The first prediction was that the duration of the syllable preceding the target vowel (independent of intonation) should have an effect on categorization, whereby overall a LONG precursor decreases "code" responses in the same way that LONG precursor decreased /p/ responses in Experiment 1. Unlike VOT, vowel duration does not robustly increase as a function of initial strengthening, in an IP-initial CV syllable (Cho & Keating, 2009) and so listeners' perception of vowel duration would not be expected to shift on the basis of their interpretation the target vowel being in an IP-initial syllable. In contrast listeners *would* be expected to modulate categorization of vowel duration (as a cue to obstruent voicing) as a function of speech rate (e.g. Heffner, Newman, & Idsardi, 2017; Saltzman, 2016). In this sense Experiment 2 is designed to test directly for speech rate effects.

Secondly, it was predicted that there will be no effect of intonation in the LONG condition. LH intonation in the SHORT condition, if it is indeed interpreted as an increase in speech rate, should affect subsequent categorization in analogous fashion to Experiment 1. Specifically, overall shorter vowel durations should be needed for a "code" response because it is perceived that the speaker has increased their speaking rate. The only difference between this experiment and Experiment 1 is that the perceived duration of the target *vowel* (instead

of the duration of VOT) is being modulated by the intentionally-driven rate percept. This predicts that in the SHORT condition only, LH intonation should *increase* "code" responses.

Experiment 2 thus allowed for further confirmation of the effect in testing if the observed outcome will extend to listeners' percepts of another durational contrast. An outcome that fits with the predictions laid out above would confirm that intonation is indeed relevant in giving listeners the impression of increased speech rate, which modulates subsequent categorization; and that the effect is generalizable to other durational contrasts.

3.1. Materials

The carrier phrase used in Experiment 2 was "I'll say coat/-code now". The words "I'll say" used in the stimuli were taken directly from Experiment 1, having two intonation conditions crossed with two duration conditions, the creation of which is laid out in Section 2.1. The words "coat/code now" were cross spliced from another sentence produced by the same speaker recorded under the same conditions as in Experiment 1. The starting point was "I'll say code now", with the same placement of pitch accent on the target (H^*) and the same L-L% boundary tone as in Experiment 1. A ToBI-transcribed representation is shown below in (5). Note the word "now" was used in lieu of "again" because it was judged to be more natural for a "coat" percept. An unreleased /t/ was judged to sound unnatural before a vowel.

(5)	...	code	now
		H^*	L-L%

The creation of the vowel duration continuum was carried out with PSOLA resynthesis (as in Experiment 1). First, audible stop voicing was removed to render the stop ambiguous. The closure duration between the coda of the target word and the following /n/ in "now" was set to be 90 ms, a relatively ambiguous value, based on previous production studies (Flege, Munro, & Skelton, 1992; Fullana & Mora, 2009). This was done so that closure duration should not bias responses, given that it is typically longer in /t/ versus /d/. The duration of the vowel (defined as the beginning of periodicity following the release of /k/ until the following stop closure) was resynthesized to create a continuum ranging from 80 ms to 220 ms in 20 ms steps (for 8 steps total). These 8 steps were spliced into the context following the words "I'll say" (from Experiment 1) in all conditions. The closure duration between the end of the vowel in "say" and the beginning of the target word was set to be the same as it was across conditions as in Experiment 1.

These manipulations created 32 unique stimuli (2 intonation conditions $\times$ 2 length conditions $\times$ 8 target vowel duration steps).

3.2. Participants

54 self-reported monolingual speakers of American English with normal hearing participated in the study. All participants were students at UCLA, and received course credit for participation. Four participants were excluded using the same criteria as in Experiment 1. The results reported here are for the remaining 50 participants (38 identifying as female, 12 identifying as male).

3.3. Procedure

The procedure for Experiment 2 was identical to that for Experiment 1, with the only difference being that participants categorized the target as "coat" or "code". As with Experiment 1, the side of the screen on which the words to be categorized appeared was counterbalanced across participants. The run time was approximately 20–25 minutes.

3.4. Results

As with Experiment 1, results are discussed in terms of the model used in their evaluation. The predictors used in the model, the contrast coding of categorical fixed effects, and the determination of the random effect structure were the same as in Experiment 1, with the only difference being that target vowel duration was treated as continuous and centered at zero (as VOT was in Experiment 1).[5]

As would be expected from any such vowel duration continuum, increasing the target vowel duration significantly increased "code" responses ($B = 2.66$; $z = 20.25$; $p < 0.001$), confirming that listeners used vowel duration as cue to obstruent voicing when categorizing the continuum.

Also as expected, length showed a significant main effect whereby a long precursor significantly decreased "code" responses ($B = -0.09$; $z = -2.54$; $p = 0.011$). Unlike Experiment 1, there was no significant interaction between vowel duration and length, suggesting that the continuum was not biased at a particular end (where in Experiment 1 the lower end of the continuum was more ambiguous). This effect of length is explainable as speech rate normalization as outlined above.

The only other significant predictor in the model was an interaction between intonation and length ($B = 0.07$; $z = 2.68$; $p < 0.01$), suggesting an asymmetrical effect of intonation across the length conditions. Such an asymmetry is expected based on the Experiment 1 results, where it was predicted that intonation should exert an influence on categorization in the SHORT condition only. To further assess the nature of the interaction, comparison of contrasts with emmeans was carried out to test the effect of intonation within each length condition. This test showed that, as expected, LH intonation significantly increased "code" responses in the SHORT condition (Estimate = -0.21; z-ratio = -3.05; $p < 0.01$), while there was no effect in the LONG condition (Estimate = 0.05; z-ratio = 0.73; $p = 0.47$), mirroring the results of Experiment 1.

This asymmetry is shown in Figs. 3 and 4. Fig. 3 shows categorization in the LONG condition, where the lines are virtually overlapping along the continuum (reflecting no effect of intonation on categorization). Fig. 4 shows categorization in the SHORT condition. In contrast to the LONG condition, LH intonation is significantly increasing "code" responses in the middle

[5] Following the same model fitting procedure as Experiment 1, simplification of the random effect structure in the model excluded two random slopes: vowel duration:length: intonation, and vowel duration:intonation. The remaining random slopes in the converging model were therefore vowel duration, length, intonation, vowel duration:length, and length: intonation.

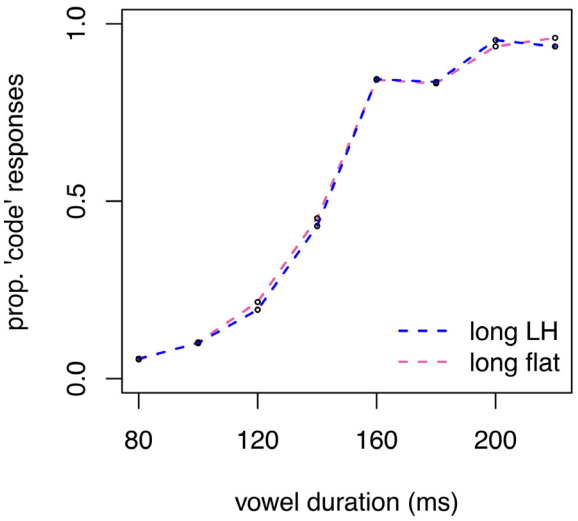

Fig. 3. Experiment 2 categorization split by intonation, in the LONG condition only.

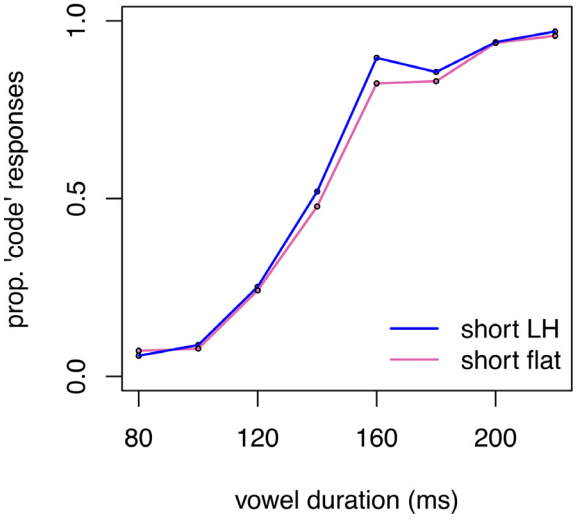

Fig. 4. Experiment 2 categorization split by intonation, in the SHORT condition only.

region of the continuum, where a clear separation can be seen. Note that the length conditions are plotted separately because the separation by length condition was not as large in magnitude as compared to Experiment 1, making the relevant asymmetry visually less clear with all conditions plotted together.

The output from the model is shown below in Table 5.

3.5. Experiment 2 discussion

The results of Experiment 2 offered confirmation of the predictions laid out at the beginning of Section 3. Specifically, LH intonation increased "code" responses only in the SHORT condition, and intonation had no effect in the LONG condition. In this way, Experiment 2 confirmed the interpretation forwarded for the results in Experiment 1, showing that categorization of vowel duration is subject to the same shift as categorization of VOT. In other words, both contrasts showed the expected effect of speech rate, informed by intonational structure. Experiment 2 also extends the results of Experiment 1 in showing that the effect persists at a greater temporal distance from

Table 5
The model output for Experiment 2. Estimates are rounded to two decimal places. Approximate *p* values are given at right.

	B (SE)	z value	*p* value
(Intercept)	0.52 (0.08)	6.55	<0.001
vdur	2.66 (0.13)	20.25	<0.001
intonation	−0.04 (0.02)	−1.68	0.09
length	−0.09 (0.03)	−2.54	0.011
vdur:intonation	−0.04 (0.03)	−1.31	0.19
vdur:length	−0.02 (0.04)	−0.41	0.68
intonation:length	0.07 (0.02)	2.68	0.007
vdur:intonation:length	−0.05 (0.03)	1.36	0.18

the target, where [k^h] intervenes between the target vowel and the precursor vowel, unlike Experiment 1 where the VOT values that were categorized immediately followed the precursor vowel [eɪ]. The results of Experiment 2 suggest that the effect of intonation in the computation of speech rate can be generalized to other comparable temporal contrasts. Further replication might test for the same effect in listeners' perception of transition duration for a /b-w/ continuum (e.g. Miller et al., 1984), for example. Taking the results from both experiments together, a picture emerges in which listeners' perception of speech rate integrates expectations about intonational structure.

However, one could argue that the observed influence of pitch on perceived speech rate may not necessarily reflect listeners' interpretation of intonation, but rather psychoacoustic perceptual differences between dynamic (rising or falling) and flat f0 contours. Previous studies have shown that dynamic contours are perceived as longer than flat contours, with duration held constant (e.g. Cumming, 2011; Lehiste, 1976), though the evidence for this claim is mixed, with some studies arguing that this occurs only in isolated monosyllables, not running speech (e.g. Van Dommelen, 1993). Nevertheless, such a possibility can be entertained in light of the present results, as the LH condition is dynamic, and therefore might be perceived as relatively long. However, this explanation is ruled out on two counts. First, if this were the case, the effect would be expected to be uniform across length conditions, which it is not. Second, if the dynamic f0 in the LH condition caused the syllable to be perceived as longer, categorization should shift in the opposite of the observed direction (e.g. longer VOT values for a /p/ response would be required following perceived lengthening). On this basis, it can be argued that the phonetic properties of the LH pitch contour alone are not responsible for the shift in categorization. Crucially, it is the function of pitch shape and tone-syllable alignment within the intonational system of the language that modulates listeners' perception of rate and subsequent categorization of durational contrasts.

One remaining question is how listeners may be interpreting the contours they hear. Specifically, is there evidence that listeners indeed perceive the SHORT LH condition as a compressed boundary tone? If they do *not* perceive the LH condition as an L-H% boundary tone, a pertinent question is what other possible interpretations they may have. Specifically, we can ask: could the low target in the SHORT LH condition be interpreted as a L^* pitch accent and the following high target as an interpolation between this L^* and the following H^*? As described in Section 2.1, the boundary in the SHORT LH condi-

tion was cued by pitch alone, lacking other cues such as decreased intensity, longer duration, or phrasal creak. Additionally, though the syllable "say" in the SHORT condition was not produced as prominent in the original stimuli (i.e. it was produced with lower intensity and shorter duration than the adjacent syllables by a ToBI-trained speaker), it is possible that the syllable was perceived as relatively prominent by listeners due to the design of the experiment itself. During the experiment, listeners heard a randomized mix of "say" in both LONG and SHORT conditions. The vowel in "say" in the SHORT condition was taken from example (3) (i.e., [I'll say pa again]$_{IP}$), and in the LONG condition it was taken from example (4) (i.e., [I'll say]$_{IP}$[pa again] $_{IP}$) above (see Section 2.1). As shown in Fig. 1, "say" in (3) was shorter in duration and stronger in amplitude than "say" in (4). Thus, the increased amplitude of the SHORT "say" may well be perceived as prominent in comparison to the relatively quiet, LONG "say". Furthermore, having a clear low f0 target during the syllable "say" in the SHORT LH condition could be perceived as L* accented, while the sagging f0 during "say" in the SHORT FLAT condition would not be perceived as prominent. In sum, the relatively stronger amplitude in the SHORT condition, as well as the tonal movement in the LH condition, may give listeners a percept of prominence on the word "say" in the SHORT LH condition, with the low f0 target interpreted as an L* pitch accent.[6] This proposal will be referred to as the *L* account*, contrasted with the *L-H% account* forwarded in preceding sections. A ToBI-transcribed representation of this hypothesized L* interpretation (for the SHORT LH condition) is shown in example (6).

(6)	I'll	say	*pa*	again
	H*	L*	H*	L-L%

The results of both Experiments 1 and 2 would be consistent with this hypothesized L* interpretation, since accented syllables are also lengthened relative to unaccented syllables (e.g. Beckman & Edwards, 1990; Cho, 2002; Turk & Sawusch, 1997; Turk & White, 1999). Therefore, a perceived pitch accent, cued by pitch movement, without lengthening, could plausibly be interpreted by listeners as an increase in speech rate, which would in turn modulates subsequent categorization, as observed in both experiments. In this sense, listeners' interpretation of an L* pitch accent, in lieu of a compressed L-H% boundary tone, might be responsible for the observed effect.

Crucially, both the L-H% account and the L* account attribute the effect observed in Experiments 1 and 2 to listener computation of speech rate on the basis of intonational structure. In this sense both are compatible with the central argument made in this paper: that intonational systems mediate speech rate normalization. The difference between the accounts is which aspect of intonational structure is responsible for the perceived increase in rate. By the L-H% account, it is assumed that listeners perceive a compressed boundary tone, and by the L* account they are assumed to perceive a pitch accent. Because both phrase-final syllables and pitch-accented syllables are lengthened, they may both generate the observed effect in Experiments 1 and 2. Further study is needed to investigate how these different aspects of prosodic structure influence listener's computation of speech rate and consequent perception of speech sounds. Testing, for example, how naturally produced pitch accents and boundaries shift categorization when compressed may provide another fruitful avenue of exploration and help inform our understanding of how the perception of speech rate is intertwined with prosody more broadly. As a preliminary investigation of this question, a third experiment testing if listeners perceive a prosodic boundary after "say" in the SHORT LH condition is contained in Appendix A.

[6] I am grateful to an anonymous reviewer for pointing out this possibility and outlining its relevance to the experimental findings.

4. General discussion

The present experiments provide evidence that listeners are sensitive to intonational structure in computing speech rate, and that these effects extend to the perception of durational (segmental) contrasts. Whether the rising pitch contour in the SHORT LH condition is interpreted as a boundary tone (L-H%) or a pitch accent (L*), a perceived increase in speech rate modulated subsequent categorization of VOT (Experiment 1) and vowel duration (Experiment 2).

As discussed in Section 1, Mitterer et al. showed that listeners adjusted categorization of VOT in the absence of tonal cues to boundary, showing that duration alone is sufficient for listeners to shift categorization. As the authors outlined, the persistent effect of duration in the absence of intonational cues could be taken to support speech rate normalization as the driving force behind their findings. The present study can compliment this view by showing that tonal cues are indeed relevant under the right circumstances. These results thus further develop the conclusions reached by Mitterer et al. in showing that speech rate is central in shifting categorization, but intonation and prosodic structure are in fact crucially intertwined with listeners' perception of rate.

These results can further be considered in light of Kim and Cho's (2013) original findings. Under the analysis that listeners interpret the SHORT LH condition as cueing an L* pitch accent, these results do not speak directly to theirs, since the perception of a (temporally compressed) pitch accent would not be interpreted as a boundary, and therefore *not* trigger the perceptual compensation for initial strengthening they discuss. Therefore, under this account, the present results could be seen as highlighting listener sensitivity to prosodic structure in a different domain than that discussed by Kim and Cho, though both have the central claim that prosody is relevant in listeners' perception of segmental contrasts. Following the L-H% account, the results *can* be compared with Kim and Cho's findings in that they show a shift in categorization with the opposite directionality as would be expected based on compensation for initial strengthening (discussed in Section 2.5). From this angle, Experiment 1 shows that when listeners are presented with a tonal cue to boundary, it appears to be most relevant in the perception of rate, in lieu of triggering compensation for initial strengthening. In this sense the results from Experiment 1 build on Kim and Cho's findings in showing that compensation for initial strengthening may be overridden in certain circumstances when it is in competition with other perceptual processes (e.g., the computation of speech rate).

Accordingly, the present results do not necessarily rule out the possibility that listeners do compensate for initial strengthening as originally argued by Kim and Cho. Further investigation of perceptual compensation for initial strengthening might circumvent the confounds discussed by Mitterer et al. in testing if and how non-durational cues are compensated for by listeners, though as outlined by Mitterer et al., initial strengthening appears to affect durational cues more than spectral ones (though see Cho & Keating, 2009; Georgeton et al., 2016; Georgeton & Fougeron, 2014). Looking more broadly at other patterns associated with prosody and intonation (e.g. phrase-final lengthening) remains promising for further addressing the general question of how the prosodic/intonational systems of a language modulate the perception of phonetic categories independently from rate-normalization.

The results in the present study also make several concrete predictions for further study. Because intonational systems are language-specific, the effect of tonal contours in cueing speech rate and prosodic structure would be expected to vary across languages with different intonational systems and categories. One promising test case is Seoul Korean, where a rising pitch pattern similar in shape to English L-H% can occur at the end of an accentual phrase (AP) without any lengthening (e.g. Jun, 1993, 1995, 1998, 2005). In fact, Kim, Mitterer, and Cho (2018) recently showed that an AP-final tonal pattern alone modulates phonological inferencing in an eye-tracking study. These results suggest that listeners are capable of computing prosodic structure independent of temporal changes, and that this computation can influence speech perception, and speech processing more generally (following e.g. Kim & Cho, 2009; Cho, McQueen, and Cox (2007). Given that listeners in Kim et al. (2018) used only tonal cues to compute prosodic structure without duration, it stands to reason they might adjust categorization of post-AP segments (with, for example, a VOT continuum) on the basis of tonal cues alone. Crucially, because the AP-final tonal contour does not robustly co-occur with lengthening, as compared to an English IP-final contour (L-H%), it would not indicate an increase in speech rate when distributed over a short vowel. Accordingly, speakers of Seoul Korean would not be predicted to shift categorization along those lines. Instead, if prosodic structure is computed without a perceived modulation in rate, it would be predicted that in an experiment similar to those in the present study, in Korean, an intonation-based shift in categorization might show the opposite directionality as the present results (reflecting compensation for initial strengthening on the basis of tonal information). Observing this sort of cross-linguistic difference would be a strong argument for the influence of language-specific intonational categories (and their temporal properties) as mediating influences in segmental speech perception. This sort of cross-linguistic research is therefore another important step in exploring how intonational systems affect the perception of fine-grained phonetic detail via speech rate normalization, the computation of prosodic structure, and their interaction.

A related theoretical implication from the present findings is that speech rate normalization, which can be seen as being early-stage, domain-general speech processing (as outlined by Mitterer et al.), is mediated to some degree by listeners' knowledge of intonation. This may seem contradictory. However, as mentioned in Section 1, the extent to which speech rate normalization is *exclusively* a reflection of domain general auditory mechanisms is an open question. Some evidence suggesting other factors are relevant can be noted. For example, lexical rate effects (e.g Dilley & Pitt, 2010), whereby the perception of whole words varies with speech rate, are influenced by language experience (e.g. Baese-Berk et al., 2016; Dilley et al., 2013). These effects also only occur when the precursor is intelligible speech, suggesting rate dependent speech processing can be speech-specific in some cases (Pitt et al., 2016).

Previous studies of speech rate effects that play out in phonetic categorization (of the kind presented here) have also shown that linguistic factors play a role. For example, Bosker and Reinisch (2015, 2017) show that perception of contrastive vowel length is influenced by language experience, where non-native speech is perceived as being spoken more quickly, modulating categorization, even when actual rate is held constant. Pind (1998) further showed that manipulating vowel spectra affects the perception of voice offset time (another temporal contrast), based on co-occurrence restrictions for certain vowels and preaspiration in Icelandic. Here again, shifts occur even with duration held constant. Purely duration-based, domain-general auditory processes clearly cannot account for cases such as these where the perception of a temporal contrast shifts even when contextual duration remains the same. These are thus empirical cases which suggest that some additional (putatively language-specific) influences are at play in the perception of temporal contrasts. The present studies can be seen as evidence that intonational systems should be considered as a mediating factor in this light as well.

Results showing linguistic influences in rate dependent speech perception have led to speculation that "different rate effects operate at different processing levels" (Bosker, 2017, p. 340), where some are domain-general and some are not. Wade and Holt (2005), for example, suggest that language experience likely plays a role in the perception of temporal contrasts, stating: "It seems likely [. . .] that speakers compensate for [. . .] variability by learning the rate-dependent covariance patterns" in the speech signal (p 948). It is also worth noting that language experience has been shown to influence the perception of non-speech (e.g. Iverson, Wagner, & Rosen, 2016; Xu, Gandour, & Francis, 2006), and even the early-stage neural representation of pitch in the auditory brain stem (Krishnan, Gandour, & Cariani, 2005, Krishnan, Gandour, Bidelman, & Swaminathan, 2009, Krishnan, Swaminathan, & Gandour, 2009). Iverson et al. suggest these sorts of results question "whether a sharp division of general auditory and speech-specific processing modes exists" in the first place (p. 1808; see also Liberman & Mattingly, 1989). This view is in accordance with the position that intonational patterns might be relevant in "low-level" perceptual processes, including the computation of speech rate.

The idea that rate-dependent speech perception is, to some extent, influenced by learned patterns in the language input fits with the present results. However, these claims remain somewhat speculative, and the relationship between general auditory processing and learned patterns is not yet well understood (e.g. Bosker, 2017; Wade & Holt, 2005). Accordingly, in going forward, cross-linguistic research of the kind

mentioned above will be key in teasing apart the effects of language experience from more domain-general auditory processes. Further studies might also benefit from the use of non-speech analogs in stimuli as another possible way of testing for the domain generality of the observed effect (following Pitt et al., 2016).

5. Conclusions

The present study consisted of two experiments, investigating the role that modulations in pitch and duration played in categorization of temporal contrasts. In Experiment 1, listeners categorized a VOT continuum, and results led to the hypothesis that listeners' sensitivity to tonal distributions in the intonational phonology of English caused them to interpret a compressed boundary tone, or pitch accent, as an increase in speech rate, shifting subsequent categorization of VOT. Experiment 2 offered support for this hypothesis, by showing an analogous shift occurred in the categorization of a vowel duration continuum, where vowel duration served as a cue to coda stop voicing. These results together are taken to suggest that tonal distributions over time are relevant to listeners' perception of speech rate, as informed by the intonational structure of their language.

The results of the present study thus suggest that we must consider intonation and prosodic structure in accounting for rate dependent speech perception. In showing that categorization is influenced by intonational patterns, the current findings indicate that prosodic/intonational structure is indeed relevant in the perception of durational contrasts. These results also support the view that speech rate normalization is central in driving the observed effect, but they point to intonational structure as a crucial meditating factor in listener computation of speech rate.

Extending these results to cross-linguistic study and exploring their relevance for more general issues in the speech perception literature will address the predictions and implications outlined above. Further research will hopefully improve our understanding of how prosodic/intonational systems, and linguistic experience more broadly, mediate rate dependent speech processing and modulate the perception of durational contrasts.

Acknowledgements

I would like to thank Sun-Ah Jun, Pat Keating, and Megha Sundara for valuable advice, as well as members of the UCLA Phonetics Lab and three anonymous reviewers for insightful comments and feedback. Further thanks to Yang Wang for assistance with data collection and to Adam Royer for recording speech for the stimuli. This research was supported in part by a UCLA Summer Research Mentorship Award (2017) to the author.

Appendix A

A third experiment outlined in this appendix was designed to provide a basic exploration of listeners' perception of a boundary in the stimuli from Experiments 1 and 2. If listeners were to interpret the SHORT LH condition as a compressed boundary tone (L-H%), they would be expected to perceive a larger prosodic juncture following the word "say" in the SHORT LH condition than that in the SHORT FLAT condition. Following this logic, Experiment 3 presented listeners with pairs of stimuli from the previous two experiments and asked them to judge in which stimulus they heard a larger break or separation between the word "say" and the word that followed in a 2AFC task (see e.g. Reinisch, 2016 for a similar approach with explicit speech rate judgments). Listeners always compared SHORT LH to SHORT FLAT stimuli (Stimuli from the SHORT condition only are used in this task, as this condition is where the crucial question of listener interpretation arises).

Materials

Since perception of the target word from the previous experiments is not of interest in this task, the two longest endpoints from each continuum were chosen to be used in this experiment (i.e. 45 ms VOT from Experiment 1, and 220 ms vowel duration from Experiment 2, corresponding to representative tokens of "pa" and "code", respectively). Therefore, there were four unique stimuli used in Experiment 3, the SHORT FLAT and SHORT LH tokens with 45 ms VOT from Experiment 1, and the SHORT FLAT and SHORT LH tokens with 220 ms vowel duration from Experiment 2.

Participants

32 self-reported monolingual speakers of American English with normal hearing participated in the study (20 identifying as female, 12 identifying as male). All participants were students at UCLA, and received course credit for participation. No participants were excluded from analysis.

Procedure

The testing location, experimental platform, and aural presentation of stimuli were the same as in Experiments 1 and 2. Prior to the experimental trials, participants were instructed that their task would be to judge how much of a separation between words they perceive, and that they should focus on how "separated" or "close together" particular words sound. They were also introduced to the term "break", which was defined as the amount of separation perceived between two words, and told that their task would be to respond to questions asking "in which sentence do you hear a *larger break*?" Participants were then told that their task was to listen to a pair of sentences and select in which sentence they perceived more of a separation/break between the word "say" and the following word. They were further given instructions about the experimental procedure (described below). There were no training trials, and no trials were excluded from analysis.

In each trial, participants were presented with two buttons visually on the computer screen, which they used to play the stimuli by clicking the button with the desktop mouse. One button was labeled "play sentence 1", and the other labeled "play sentence 2". To listen to the stimuli, participants clicked the labeled buttons, and were allowed to listen to each stimulus a maximum of two times. After listening to the stimuli, participants selected in which stimulus they perceived a larger break between "say" and the following word by pressing one of the two response buttons presented below the playback buttons. One button was labeled "sentence 1" and the other labeled "sentence 2". In every trial, one stimulus had SHORT LH intonation, and the other had SHORT FLAT intonation. The inter-trial-interval was 250 ms. The experiment consisted of 48 trials.

24 trials contained comparisons of SHORT LH and SHORT FLAT stimuli from Experiment 1, and the other 24 contained comparisons of SHORT LH and SHORT FLAT stimuli from Experiment 2. The trials were blocked by which experiment the stimuli originated from. The order of presentation of the two blocks was counter balanced across participants (16 heard the Experiment 1 block first, 16 heard the Experiment 2 block first).

The crucial manipulation is therefore whether the SHORT LH token is sentence 1, or sentence 2 (where the SHORT FLAT token is always the other sentence). The SHORT LH token was sentence 1 in half of all trials in the experiment. This variable was randomized within block, meaning listeners heard a randomized mix of SHORT LH and SHORT FLAT tokens as sentence 1. Given this set up, the crucial prediction is that if listeners interpret the tonal movement present in the SHORT LH condition as cuing a boundary, they would select the sentence with SHORT LH intonation as having a larger break between "say" and the following word more often than the SHORT FLAT condition. In other words, participants' response ("sentence 1" or "sentence 2") should be significantly influenced by whether a sentence has LH or FLAT intonation.

Results and discussion

The results from Experiment 3 were assessed by a linear mixed-effect model with a logistic linking function, predicting participants' choice of sentence ("sentence 1" or "sentence 2"). Results are plotted and reported such that the dependent variable is the response "sentence 1 has a larger juncture" ("sentence 1" coded as 1, "sentence 2" coded as 0). The crucial predictor is the ordering of intonation conditions, that is, whether LH or FLAT intonation was in sentence 1 (where the other stimulus was always sentence 2). This was specified as a fixed effect in the model and will be referred to as "sentence 1 intonation", with two levels, FLAT and LH. As in previous experiments, this variable was effect-coded (in this experiment, FLAT was mapped to −1, while LH was mapped to 1). The model was specified with by-subject and by-item random intercepts, where item refers to the sentence in which the target word "say" was embedded (either "I'll say pa again" from Experiment 1, or "I'll say code now" from Experiment 2). The model was also fit with the maximal number of by-subject and by-item random slopes. The fully specified model converged, meaning the random effect structure was not simplified.

Given this coding scheme, a positive estimate for "sentence 1 intonation" would show that "sentence 1" responses increase when sentence 1 has LH intonation (and sentence 2 has FLAT intonation), showing that listeners tend to select a sentence with LH intonation as having a larger break. A positive coefficient, and a significant effect of "sentence 1 intonation" is therefore predicted if listeners perceive a larger juncture in the SHORT LH stimuli. An effect plot for the model, showing "sentence 1" responses as a function of sentence 1 intonation (LH vs FLAT) is given in Fig. 5.

As visualized in Fig. 5, the intonation of sentence 1 significantly influenced whether listeners perceived a larger juncture in that sentence. Specifically, listeners were more likely to perceive sentence 1 as having a larger juncture when sentence 1 had LH intonation. In the model this is reflected by a positive coefficient for "sentence 1 intonation", and a significant effect ($B = 0.33$; $z = 2.70$; $p < 0.01$). In general terms, this indicates

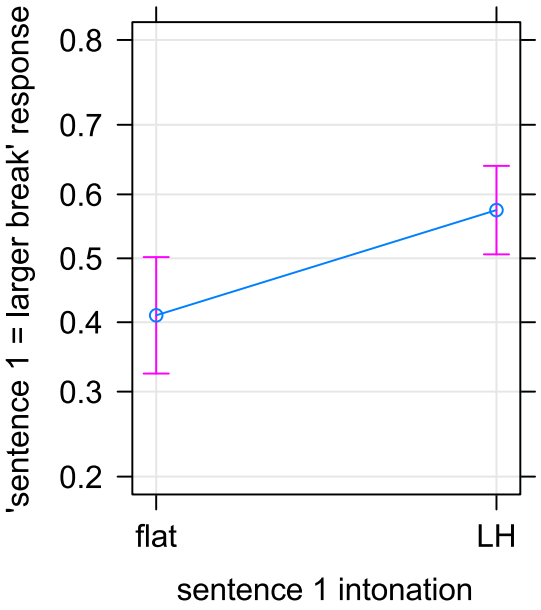

Fig. 5. The effect of sentence 1 intonation on participant responses in Experiment 3. The points show the model fit for a "sentence 1" response, for each sentence 1 intonation condition (shown on the on the *x* axis). Error bars show one standard error. The plot was generated using the 'effects' package in R (Fox, 2003).

Table 6
The model output for Experiment 3. Estimates are rounded to two decimal places. Approximate *p* values are given at right.

	B (SE)	z value	*p* value
(Intercept)	−0.03 (0.11)	−0.25	0.80
sentence 1 int.	0.33 (0.12)	2.70	0.007

that listeners perceived a larger break in the LH condition than in the FLAT condition. The model output is shown in Table 6.

These results may be taken as broadly supportive of the idea that the tonal movement on "say" in Experiments 1 and 2 is interpreted by listeners as signaling a boundary. This interpretation can explain both the shift in categorization found in Experiments 1 and 2, as well as the larger perceived juncture found in Experiment 3. However, it can be seen the participants' responses are far from categorical. Though the group data shows a robust effect with significantly more "larger break" responses in the LH condition, some listeners selected LH intonation as a larger juncture near chance. Other listeners showed a clear preference, selecting LH intonation as a larger juncture in 75–90% of trials, in line with the observed effect in the model. This individual variability leaves open the possibility that different listeners interpreted the LH contour differently. Notably, no participant selected the FLAT contour as a larger perceived break reliably. This seems to indicate clearly that no boundary is perceived by listeners in the FLAT condition, as expected.

The observed listener variability and the small size of the effect observed in Experiment 3 may also shed light on the relatively small effect observed in Experiments 1 and 2. Given that the SHORT LH condition is perceived as a larger break only approximately 58% of the time, perception of an increase in speech rate may simply not occur for listeners who do have this interpretation (though an L^* interpretation is not tested here). The limited scope of the effect in Experiment 3 may

therefore shed light on the relatively small magnitude of the effect of intonation in the SHORT condition observed in Experiments 1 and 2, where no shift in categorization would be expected to occur in cases where listeners are not interpreting the stimuli as predicted.

The results of Experiment 3 also do not preclude the possibility that listeners interpret the SHORT LH condition as prominence-lending, both in light of the limited nature of the effect, and given that tendency to perceive an increased juncture does not necessarily rule out perception of increased prominence, accentuation etc. Therefore, both the L* and L-H % account remain viable as an explanation for the results in Experiment 1 and 2, though Experiment 3 can be taken as offering some support for perception of a boundary. As outlined in Section 3.5, both accounts crucially implicate listeners' sensitivity to intonational patterns in their computation of speech rate and categorization of speech sounds, and therefore both are compatible with the central argument forwarded above.

Appendix B

Supplementary materials. Files that exemplify the manipulations can be found in the Mendeley data repository, at https://data.mendeley.com/datasets/fcwdtc73gj/1. The files correspond to the four conditions shown in Fig. 1 and are named accordingly. All example files contain a target sound with 45 ms VOT (from Experiment 1). This is summarized in Table 7. Table 8 gives a readout of the R code used in the model for each experiment.

Table 7
Description of linked sound files.

Sound file name	Intonation condition	Duration condition
LH_short_45.wav	LH	SHORT
LH_long _45.wav	LH	LONG
flat_short _45.wav	FLAT	SHORT
flat_ long _45.wav	FLAT	LONG

Table 8
Description of model fixed and random effects, given in R syntax. Note that for Experiments 1 and 2, correlation parameters are removed such that the syntax for random effects is slightly different (see Bates, Maechler, et al., 2015, p. 7).

Experiment	Model specification
1	response ~ VOT*length*intonation + (1 | subj) + (0 + VOT | subj) + (0 + length | subj) + (0 + intonation | subj) + (0 + VOT:length | subj) + (0 + VOT:intonation | subj) + (0 + length:intonation | subj)
2	response ~ vowel_duration*length*intonation + (1 | subj) + (0 + vowel_duration | subj) + (0 + length | subj) + (0 + intonation | subj) + (0 + vowel_duration:length | subj) + (0 + length:intonation | subj)
3	response ~ sentence_1_intonation + (1 + sentence_1_intonation | subject) + (1 + sentence_1_intonation | word)

Reference

Baese-Berk, M. M., Morrill, T. H., & Dilley, L. C. (2016). Do non-native speakers use context speaking rate in spoken word recognition? Proceedings of the 8th International Conference on Speech Prosody (SP2016), 2016-January (pp. 979–983).

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3).

Bates, D., Kliegl, R., Vasishth, S., & Baayen, R. H. (2015b). Parsimonious mixed models. Available from arXiv:1506.04967 (stat.ME).

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015a). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48.

Bech, M., & Gyrd-Hansen, D. (2005). Effects coding in discrete choice experiments. *Health Economics, 14*(10), 1079–1083.

Beckman, M. E., & Edwards, J. (1990). Lengthenings and shortenings and the nature of prosodic constituency. In J. Kingston & M. E. Beckman (Eds.), *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech*. Cambridge: Cambridge University Press.

Beckman, M. E., & Pierrehumbert, J. B. (1986). Intonational structure in Japanese and English. *Phonology, 3*(01), 255–309.

Boersma, P., & Weenik, D. (2018): Praat: doing phonetics by computer [Computer program]. Version 6.0.37, retrieved from http://www.praat.org/.

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

Bosker, H. R., & Reinisch, E. (2015). Normalization for speech rate in native and nonnative speech. Proceedings of the 18th International Congresses of Phonetic Sciences (ICPhS 2015).

Bosker, H. R., & Reinisch, E. (2017). Foreign languages sound fast: evidence from implicit rate normalization. *Frontiers in Psychology*, 8.

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

Byrd, D. (2000). Articulatory vowel lengthening and coordination at phrasal junctures. *Phonetica, 57*(1), 3–16.

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

Cho, T. (2002). *The effects of prosody on articulation in English*. New York, NY: Routledge.

Cho, T. (2015). Language effects on timing at the segmental and suprasegmental levels. In M. A. Redford (Ed.), *The handbook of speech production* (pp. 505–529). John Wiley & Sons Inc..

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

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

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

Cumming, R. (2011). The effect of dynamic fundamental frequency on the perception of duration. *Journal of Phonetics, 39*(3), 375–387.

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

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

Dilley, L., Morrill, T., & Banzina, E. (2013,). New tests of the distal speech rate effect: examining cross-linguistic generalization. *Frontiers in Psychology, 4*.

Dilley, L. C., & Pitt, M. A. (2010). Altering context speech rate can cause words to appear or disappear. *Psychological Science, 21*(11), 1664–1670.

Flege, J. E., Munro, M. J., & Skelton, L. (1992). Production of the word-final English /t/-/d/ contrast by native speakers of English, Mandarin, and Spanish. *The Journal of the Acoustical Society of America, 92*(1), 128–143.

Fougeron, C. (1998). *Variations articulatoires en début de constituents prosodiques de différents niveaux en Français* (Dissertation). Paris: Université Paris III Sorbonne Nouvelle.

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

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

Fox, J. (2003). Effect displays in R for generalised linear models. *Journal of Statistical Software, 8*(15), 1–27.

Fullana, N., & Mora, J. C. (2009). Production and perception of voicing contrasts in english word-final obstruents: Assessing the effects of experience and starting age. In M. A. Watkins, A. S. Rauber, & B. O. Baptista (Eds.), *Recent research in second language phonetics/phonology: perception and production* (pp. 97–117). Newcastle upon Tyne, UK: Cambridge Scholars Publishing.

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

Georgeton, L., & Fougeron, C. (2014). Domain-initial strengthening on French vowels and phonological contrasts: Evidence from lip articulation and spectral variation. *Journal of Phonetics, 44*, 83–95.

Harrington, J., Kleber, F., & Reubold, U. (2008). Compensation for coarticulation, /u/-fronting, and sound change in standard southern British: An acoustic and perceptual study. *The Journal of the Acoustical Society of America, 123*(5), 2825–2835.

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

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

Iverson, P., Wagner, A., & Rosen, S. (2016). Effects of language experience on pre-categorical perception: Distinguishing general from specialized processes in speech perception. *The Journal of the Acoustical Society of America, 139*(4), 1799–1809.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*(4), 434–446.
Jun, S.-A. (1993). *The phonetics and phonology of Korean Prosody* (Doctoral dissertation). The Ohio State University.
Jun, S.-A. (1995). Asymmetrical prosodic effects on the laryngeal gesture in Korean. In B. Connell & A. Arvaniti (Eds.), *Phonology and phonetic evidence: Papers in Laboratory Phonology IV* (pp. 235–253). Cambridge University Press.
Jun, S.-A. (1998). The accentual phrase in the korean prosodic hierarchy. *Phonology, 15*(2), 189–226.
Jun, S.-A. (2005). Korean intonational phonology and prosodic transcription. In S.-. A. Jun (Ed.), *Prosodic typology* (pp. 201–229). New York: Oxford University Press.
Keating, P. (2006). Phonetic encoding of prosodic structure. In J. Harrington & M. Tabain (Eds.), *Speech production: Models, phonetic processes, and techniques* (pp. 167–186). New York and Hove: Macquarie Monographs in Cognitive Science, Psychology Press.
Keating, P., Fougeron, C., Hsu, C., & Cho, T. (2003). Domain initial articulatory strengthening in four languages. In J. Local, R. Ogden, & R. Temple (Eds.), *Phonetic interpretation: Papers in Laboratory Phonology VI*. Cambridge University Press.
Kim, S. (2004). *The role of prosodic phrasing in Korean word segmentation* (Doctoral dissertation). UCLA.
Kim, S., & Cho, T. (2009). The use of phrase-level prosodic information in lexical segmentation: evidence from word-spotting experiments in Korean. *The Journal of the Acoustical Society of America, 125*(5), 3373–3386.
Kim, S., & Cho, T. (2013). Prosodic boundary information modulates phonetic categorization. *The Journal of the Acoustical Society of America, 134*(1), EL19–EL25.
Kim, S., Mitterer, H., & Cho, T. (2018). A time course of prosodic modulation in phonological inferencing: The case of Korean post-obstruent tensing. *PLoS ONE, 13*(8) e0202912.
Kjelgaard, M. M., & Speer, S. R. (1999). Prosodic facilitation and interference in the resolution of temporary syntactic closure ambiguity. *Journal of Memory and Language, 40*, 153–194.
Klatt, D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *The Journal of the Acoustical Society of America, 59*(5), 1208–1221.
Krishnan, A., Gandour, J. T., Bidelman, G. M., & Swaminathan, J. (2009a). Experience dependent neural representation of dynamic pitch in the brainstem. *Neuroreport, 20*(4), 408–413.
Krishnan, A., Swaminathan, J., & Gandour, J. T. (2009b). Experience-dependent enhancement of linguistic pitch representation in the brainstem is not specific to a speech context. *Journal of Cognitive Neuroscience, 21*(6), 1092–1105.
Krishnan, A., Xu, Y., Gandour, J., & Cariani, P. (2005). Encoding of pitch in the human brainstem is sensitive to language experience. *Brain Research. Cognitive Brain Research, 25*(1), 161–168.
Ladd, D. R. (2008). *Intonational phonology*. Cambridge University Press.
Ladd, D. R., & Schepman, A. (2003). "Sagging transitions" between high pitch accents in English: Experimental evidence. *Journal of Phonetics, 31*(1), 81–112.
Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *The Journal of the Acoustical Society of America, 29*(1), 98–104.
Lee, E.-K., & Watson, D. G. (2011). Effects of pitch accents in attachment ambiguity resolution. *Language and Cognitive Processes, 26*(2), 262–297.
Lehiste, I. (1976). Influence of fundamental frequency pattern on the perception of duration. *Journal of Phonetics, 4*(2), 113–117.
Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2018). emmeans: Estimated marginal means, aka least-squares means. Retrieved from: https://CRAN.R-project.org/package=emmeans.
Liberman, A. M., & Mattingly, I. G. (1989). A specialization for speech perception. *Science, 243*(4890), 489–494.
Lisker, L. (1986). "Voicing" in English: a catalogue of acoustic features signaling /b/ VERSUS /p/ in trochees. *Language and Speech, 29*(1), 3–11.
Mann, V. A. (1980). Influence of preceding liquid on stop-consonant perception. *Perception & Psychophysics, 28*(5), 407–412.
Mann, V. A., & Repp, B. H. (1981). Influence of preceding fricative on stop consonant perception. *The Journal of the Acoustical Society of America, 69*(2), 548–558.
Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language, 94*, 305–315.
Miller, J. L., Aibel, I. L., & Green, K. (1984). On the nature of rate-dependent processing during phonetic perception. *Perception & Psychophysics, 35*(1), 5–15.
Miller, J. L., & Volaitis, L. E. (1989). Effect of speaking rate on the perceptual structure of a phonetic category. *Perception & Psychophysics, 46*(6), 505–512.
Mitterer, H. (2006). On the causes of compensation for coarticulation: Evidence for phonological mediation. *Perception & Psychophysics, 68*(7), 1227–1240.
Mitterer, H., Cho, T., & Kim, S. (2016). How does prosody influence speech categorization? *Journal of Phonetics, 54*, 68–79.
Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication, 9*(5–6), 453–467.
Newman, R. S., & Sawusch, J. R. (1996). Perceptual normalization for speaking rate: effects of temporal distance. *Perception & Psychophysics, 58*(4), 540–560.
Newman, R. S., & Sawusch, J. R. (2009). Perceptual normalization for speaking rate III: Effects of the rate of one voice on perception of another. *Journal of Phonetics, 37*(1), 46–65.
Onaka, A. (2003). Domain-initial strengthening in Japanese: An acoustic and articulatory study. In Proceedings of the 15th international congress of phonetic sciences. (pp. 2091–2094). Barcelona, Spain.
Pierrehumbert, J. (1980). *The phonology and phonetics of English intonation* (PhD). Massachusetts: Institute of Technology.
Pierrehumbert, J., & Talkin, D. (1992). Lenition of / h / and glottal stop. In G. Doherty & D. R. Ladd (Eds.), *Papers in Laboratory Phonology II: Gesture segment prosody* (pp. 90–116). Cambridge University Press.
Pind, J. (1998). Auditory and linguistic factors in the perception of voice offset time as a cue for preaspiration. *The Journal of the Acoustical Society of America, 103*(4), 2117–2127.
Pisoni, D. B., Carrell, T. D., & Gans, S. J. (1983). Perception of the duration of rapid spectrum changes in speech and nonspeech signals. *Perception & Psychophysics, 34*(4), 314–322.
Pitt, M. A., Szostak, C., & 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. *Attention, Perception, & Psychophysics, 78*(1), 334–345.
Price, P. J., Ostendorf, M., Shattuck-Hufnagel, S., & Fong, C. (1991). The use of prosody in syntactic disambiguation. *The Journal of the Acoustical Society of America, 90*(6), 2956–2970.
Raphael, L. J. (1972). Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English. *The Journal of the Acoustical Society of America, 51*(4B), 1296–1303.
Reinisch, E. (2016). Natural fast speech is perceived as faster than linearly time-compressed speech. *Attention, Perception, & Psychophysics, 78*(4), 1203–1217.
Reinisch, E., & Sjerps, M. J. (2013). The uptake of spectral and temporal cues in vowel perception is rapidly influenced by context. *Journal of Phonetics, 41*(2), 101–116.
Rietveld, A. C. M., & Gussenhoven, C. (1987). Perceived speech rate and intonation. *Journal of Phonetics, 15*(3), 273–285.
RStudio Team. (2016). *RStudio: Integrated Development for R*. Boston, MA: RStudio Inc..
Saltzman, D. (2016). *The role of the speech envelope in speaking rate compensation* (M. S.). Villanova University.
Schafer, A. (1996). Focus in relative clause construal. *Language and Cognitive Processes, 11*(1–2), 135–164.
Shattuck-Hufnagel, S., & Turk, A. E. (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research, 25*(2), 193–247.
Sjerps, M. J., & Smiljanić, R. (2013). Compensation for vocal tract characteristics across native and non-native languages. *Journal of Phonetics, 41*(3), 145–155.
Spinelli, E., Grimault, N., Meunier, F., & Welby, P. (2010). An intonational cue to word segmentation in phonemically identical sequences. *Attention, Perception, & Psychophysics, 72*(3), 775–787.
Streeter, L. A. (1978). Acoustic determinants of phrase boundary perception. *Journal of the Acoustical Society of America, 64*(6), 1582–1592.
Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology. Human Perception and Performance, 7*(5), 1074–1095.
Tehrani, H. (2015). Appsobabble [online applications platform] http://www.appsobabble.com.
Turk, A. E., & Sawusch, J. R. (1997). The domain of accentual lengthening in American English. *Journal of Phonetics, 25*(1), 25–41.
Turk, A. E., & Shattuck-Hufnagel, S. (2007). Multiple targets of phrase-final lengthening in American English words. *Journal of Phonetics, 35*(4), 445–472.
Turk, A. E., & White, L. (1999). Structural influences on accentual lengthening in English. *Journal of Phonetics, 27*(2), 171–206.
Tyler, M. D., & Cutler, A. (2009). Cross-language differences in cue use for speech segmentation. *The Journal of the Acoustical Society of America, 126*(1), 367–376.
Van Dommelen, W. (1993). Does dynamic f0 increase perceived duration? New light on an old issue. *Journal of Phonetics, 21*(4), 367–386.
Veilleux, N., Shattuck-Hufnagel, S., & Brugos, A. 6.911 Transcribing prosodic structure of spoken utterances with ToBI. January IAP 2006. Massachusetts Institute of Technology: MIT OpenCourseWare, https://ocw.mit.edu. License: Creative Commons BY-NC-SA.
Wade, T., & Holt, L. L. (2005). Perceptual effects of preceding nonspeech rate on temporal properties of speech categories. *Perception & Psychophysics, 67*(6), 939–950.
Walsh, T., & Parker, F. (1981). Vowel length and voicing in a following consonant. *Journal of Phonetics, 9*(3), 305–308.
Warner, N., Otake, T., & Arai, T. (2010). Intonational structure as a word-boundary cue in Tokyo Japanese. *Language and Speech, 53*(1), 107–131.
Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. J. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *The Journal of the Acoustical Society of America, 91*(3), 1707–1717.
Xu, Y., Gandour, J. T., & Francis, A. L. (2006). Effects of language experience and stimulus complexity on the categorical perception of pitch direction. *The Journal of the Acoustical Society of America, 120*(2), 1063–1074.