UNIVERSITY OF CALIFORNIA

Los Angeles

Intonation mediates speech rate normalization in the perception of segmental categories

A thesis submitted in partial satisfaction

of the requirements for the degree Master of Arts

in Linguistics

by

Jeremy Andrew Steffman

2018

ABSTRACT OF THE THESIS

Intonation mediates speech rate normalization

in the perception of segmental categories

by

Jeremy Andrew Steffman

Master of Arts in Linguistics

University of California, Los Angeles, 2018

Professor Sun-Ah Jun, Chair

This thesis reports on two experiments that address the question of how listeners are sensitive to prosodic/intonational structure in their perception of segmental contrasts. 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 (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.

The thesis of Jeremy Andrew Steffman is approved.

Megha Sundara

Patricia Keating

Sun-Ah Jun, Committee Chair

University of California, Los Angeles

2018

For Kathy and Laura

Table of contents

List of Tables

List of Figures

Acknowledgements

I could not have completed this thesis in its current state without the help of many. I would first like to thank Sun-Ah Jun, my chair, who provided me with invaluable guidance and encouragement throughout the process. Carrying out this project has been enormously rewarding, and this is due in no small part to her advice and support. I am also very grateful to my other committee members, Megha Sundara and Pat Keating, whose feedback and insight substantially developed the project and helped me better articulate my thinking.

Further thanks to Adam Royer for producing the speech from which the stimuli were created and to Yang Wang for help with data collection. I also would like to thank Jesse Harris for advice that was very helpful in the beginning stages of the project, as well as audience members and conference attendees at the UCLA phonetics seminar, the 174th meeting of the ASA, BLS 44, and HISPhonCog 2018 for feedback and ideas on various stages of the project.

Finally, I am grateful to my family members, for their support, and for tolerating the pilot stimuli I asked them to listen to.

1 Introduction

1.1 Background

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 et al., 2016; Jun, 1993; Onaka, 2003; Wightman et al., 1992). 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 et al., 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, Cho & Kim, 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 et al., 2008; Mann, 1980; Mann & Repp, 1981; Mitterer, 2006), formant frequency distributions preceding the target (e.g. Holt et al., 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 & 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.[1] 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 hear *x* again
H* H* L-L%

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

[1]Note Kim & Cho's pitch accent manipulation is left aside here; manipulating pitch accent placement did not have a significant effect in their experiment.

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 challenged more recently by Mitterer, Cho & Kim (2016), who present speech rate normalization as an 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 et al. (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 & 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 & Cho might be attributable to rate changes alone (for example, Miller & Volaitis found

approximately 8ms shifting; note their durational difference across rate conditions was slightly larger than Kim & 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 & Cho. Arguments for domain generality come from the fact that rate normalization can occur across changes in speaker (e.g. Diehl et al., 1980; Newman & Sawusch, 2009), and that its effects can be replicated with non-speech analogs (e.g. Diehl & Walsh, 1989; Pisoni et al., 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 et al., 2016; Bosker & Reinisch, 2015, 2017; Dilley et al., 2013), and other research indicates some rate effects are speech specific (Pitt et al., 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 4.

Because of the potentially confounding influence of speech rate normalization outlined above, Mitterer et al. performed two experiments aimed at testing Kim & Cho's original claim. In one experiment, Mitterer et al. used the same stimuli as Kim & 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 et al., 2010; Warner et al., 2010) and syntactic/phrasal grouping and disambiguation (e.g. Kjelgaard & Speer, 1999; Lee & Watson, 2011; Price et al., 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, and concluded that listeners may not be sensitive to prosodically driven variation for the purpose of categorizing segments.

1.2 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 duration cues to an IP boundary in a 2x2 design, using a similar 2AFC task as that implemented by Kim & 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 2x2 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 intonation 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 English an IP-final syllable can have four possible (non-downstepped) F0 contours: a high rise (H-H%), a high flat (H-L%), a low rise (L-H%), and a low fall (L-L%). 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, only a low-rise (L-H%) uniquely signals 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% could 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 short syllable cannot be interpreted as part of any pitch accent or a transition tone or a delayed tone. In theory, L-H% could be interpreted as an L*+H pitch accent, but this would be possible only when the syllable is pitch accented, i.e., produced prominently. Note, in the current experiment, "I'll" and "x" are pitch accented and "say" is not. Further, the low target in L-H% 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. In summary: in English intonational phonology, a non-pitch

accented syllable with two tonal targets has to be an IP-final syllable carrying a L-H% boundary tone.

Following this logic, the L-H% contour should sound like a boundary tone to 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.

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

Table 1: *Conditions used in the experiment, with predicted listener interpretation. Intonation conditions are columns, length conditions are rows.*

	INTONATION	
	Low-rising F0 (LH)	**High-flat F0 (FLAT)**
SHORT vowel	(a) IP boundary with an L-H% boundary tone (even when SHORT).	(b) Not interpretable as a boundary tone.
LONG vowel	(c) IP boundary with an L-H% boundary tone.	(d) IP boundary with an H-L% boundary tone.

As shown in Table 1 above, in the SHORT condition only, the intonational contour should sound like a boundary tone when LH, but not when FLAT. 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 below.

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.

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.

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 & Cho (2013) and Mitterer et al. (2016).

2 Experiment 1

Following Kim & 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 [pʰɑ] during recording, written as 'pa' below.

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

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

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 (b) 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 245 ms was reinserted into the duration-normalized frame to create condition (c) (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) (FLAT F0 on a LONG vowel).

Figure 1 below shows the pitch tracks and spectrograms for the four conditions.

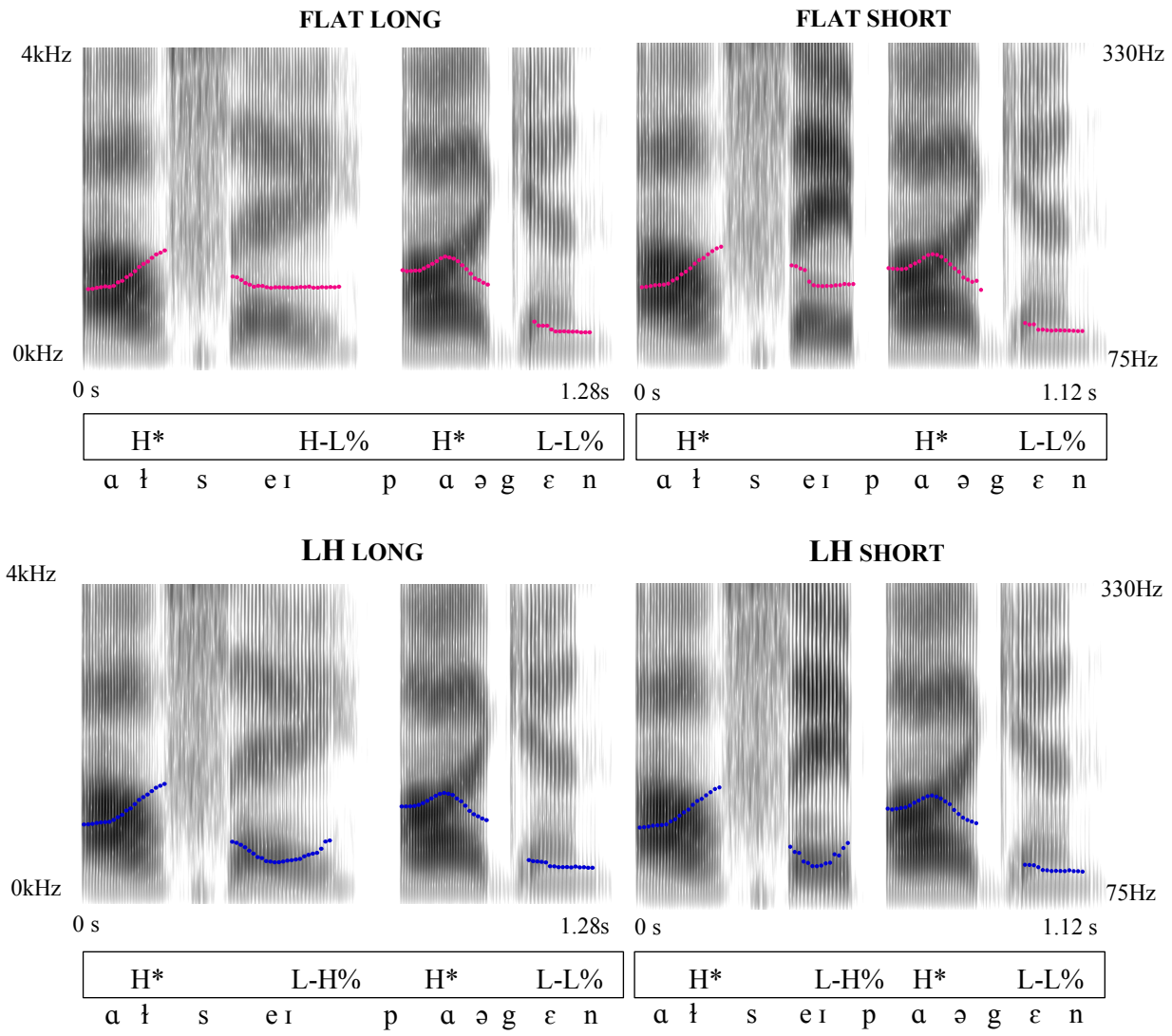

Figure 1: *Spectrograms of the stimuli sentence "I'll say pa/ba again", overlaid with pitch tracks, for the two intonation and two length conditions. ToBI transcriptions based on the predicted listener interpretations are given below each spectrogram, as well as a segmental transcription. The stimuli shown have 0ms of VOT for the target stop. Approximate times are also given below each spectrogram. The pitch range is shown at right (75-330Hz), while the frequency range for the spectrograms is shown at left (0-4kHz)*

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 milliseconds. These are similar to the values used by Kim & 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 & 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 [p^h] in (4) (with a duration of approximately 60ms), 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 x 2 length conditions x 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 (they 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 endpoints of the continuum, participants first heard the endpoints of the continuum and were told the participant was saying /bɑ/ at the 0 ms endpoint, and saying /pɑ/ at the 45 ms endpoint. 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 F0 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 ITI (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 minutes.

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 output from the model and additional information about the model selection process are included in Appendix A. 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 et al., 2015), 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), as opposed to dummy coded, 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 F0 (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 et al., 2013). Data visualization was carried out in RStudio (RStudio team, 2016). In the plot of the categorization function, the *x* axis shows the VOT values from the continuum, while the *y* axis shows the percentage of /p/ responses at each value.

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 ($B = 2.77$, $z = 17.42$, $p < 0.001$).

Length also showed a significant main effect. As expected, a LONG preceding vowel significantly decreased /p/ responses ($B = -0.24$, $z = -4.55$, $p < 0.001$), shown in Figure 2 below. Length also showed a significant interaction with VOT ($B = 0.35$, $z = 6.55$, $p < 0.001$), whereby as VOT increased the effect of length diminished along the continuum. This effect can also be observed in Figure 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.[2]

The intonation manipulations also significantly shifted categorization, where the FLAT condition significantly decreased /p/ responses ($B = -0.06$, $z = -2.24$, $p < 0.05$). 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 ($B = -0.08$, $z = -2.44$, $p = 0.015$). To further investigate the interaction, the lsmeans post hoc test was used (Lenth, 2016), where the effect of intonation in each length condition was compared at each VOT value along the continuum (see Appendix A, Table 7). 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 (0-20ms VOT), there was a significant effect of intonation ($p < 0.01$), and from 25-45ms there was no effect. This asymmetry is visible in Figure 2 below, where in the SHORT condition only, the lines of the categorization function are consistently

[2] 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 [p^h]. 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 & Cho (2013).

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.

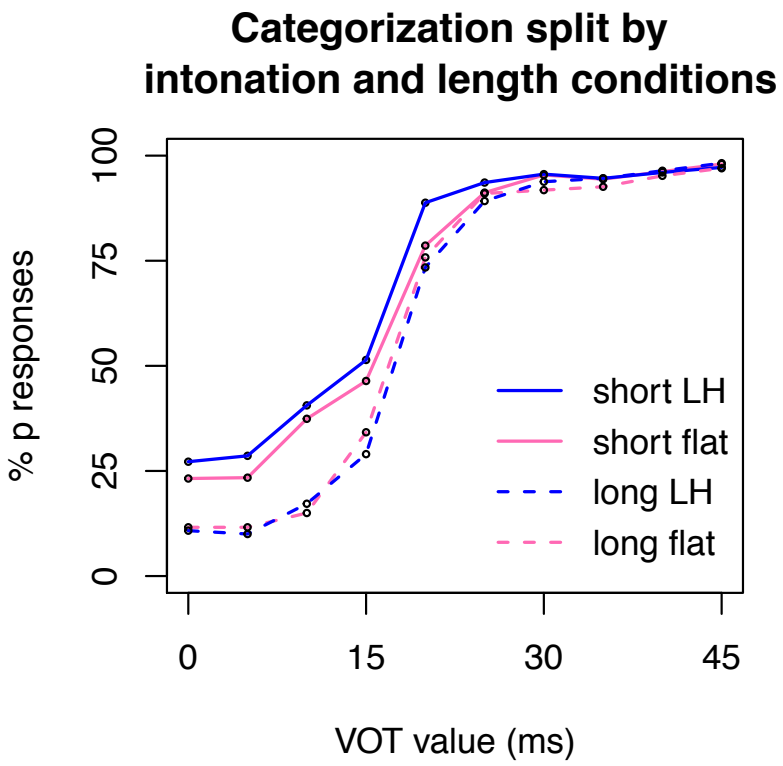

Figure 2: *Experiment 1 categorization split by all four conditions.*

2.5 Interim discussion

The results reported in 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 & Cho (2013) and Mitterer et al. (2016), where a lengthened precursor shifted the categorization to higher VOT values for a voiceless /p/ categorization. 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).

As shown in Figure 2 above, in the SHORT condition only, LH intonation *increased* /p/ responses, with a significant effect at lower, more ambiguous, VOT values. Importantly, this is the opposite of what would be expected based on Kim & Cho's original claim. Following Kim & Cho, if LH pitch in the SHORT condition indicated that the syllable preceding the target was phrase final, and if listeners used this information to modulate categorization of the VOT continuum in line with initial strengthening, it would be expected that LH pitch would *reduce* /p/ responses in the SHORT condition (see Table 2).

Instead, these unexpected results (increased, not decreased /p/ responses) suggest a novel interpretation: that listeners are using intonational information to compute speech rate. 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. 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, which altered subsequent categorization of VOT. 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.

Previous studies have suggested that dynamic F0 contours (contours involving rising or falling pitch) 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 researchers arguing that this occurs only in isolated monosyllables, not running speech (e.g. Van Dommelen, 1993). Such a possibility can be entertained in light of the present results, as the LH condition is dynamic and the FLAT condition is not. However this explanation for the present results 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 (i.e. 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. Instead, its status as a boundary tone in English intonational phonology, with restricted distribution at phrasal edges and co-occurrent lengthening, must be what drives the effect.

The suggestion that intonational structure influences how speakers perceive speech rate has in fact been made before. Rietveld & Gussenhoven (1987) show 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 that differ from the segmental categorization task used here. Accordingly, their findings are broadly compatible with the present results, though they do not offer direct support for the current 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 2x2 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 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. 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 intonationally-driven rate percept. In terms of the statistical model, 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).[3]

(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 et al., 1991; 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 x 2 length conditions x 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] 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.

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). The model output and more information regarding the model selection process is contained in Appendix B.

As would be expected from any such vowel duration continuum, increasing the target vowel duration significantly increased “code” responses (B = 2.67, z = 20.33, $p < 0.001$), 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 = -0.10, z = -2.78, $p < 0.01$). 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 only as speech rate normalization as outlined above.

The only other significant predictor in the model was an interaction between intonation and length (B = 0.06, z = 2.71, 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, the least-squared means post hoc test was used 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 = -3.07, p < 0.01), while there was no effect in the LONG condition (Estimate = -0.05, z-ratio = 0.3, p = 0. 46), mirroring the results of Experiment 1. This asymmetry is shown in Figures 3 and 4. Figure 3 shows categorization in the LONG condition, where the lines are virtually overlapping along the continuum (reflecting no effect of intonation on categorization).

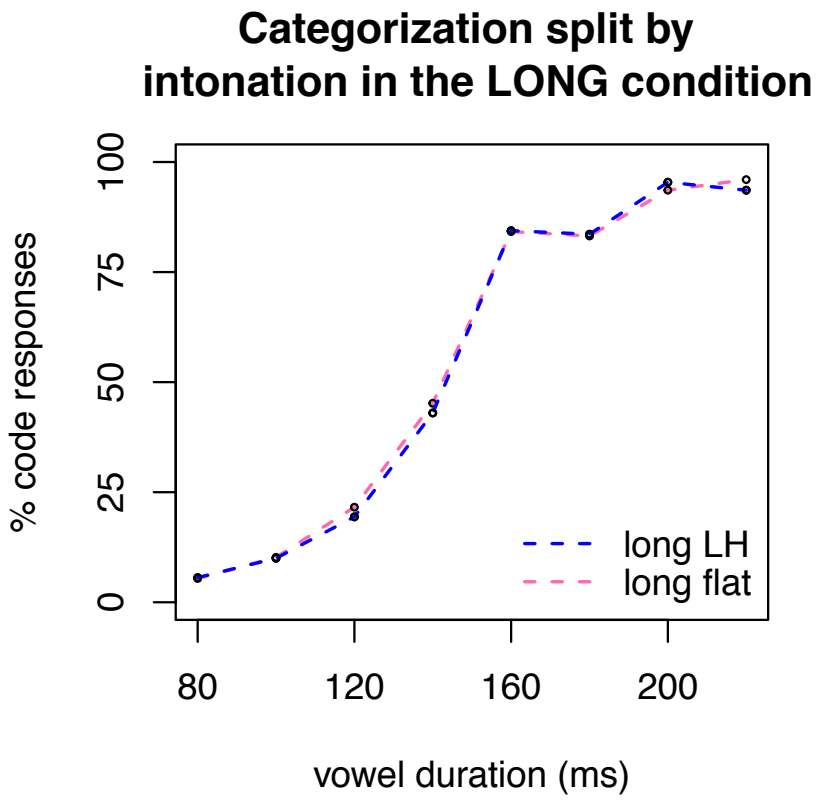

Figure 3: *Experiment 2 categorization split by intonation, in the* LONG *condition.*

Figure 4 shows categorization in the SHORT condition. In contrast to the LONG condition, LH intonation is significantly increasing "code" responses in the middle 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.

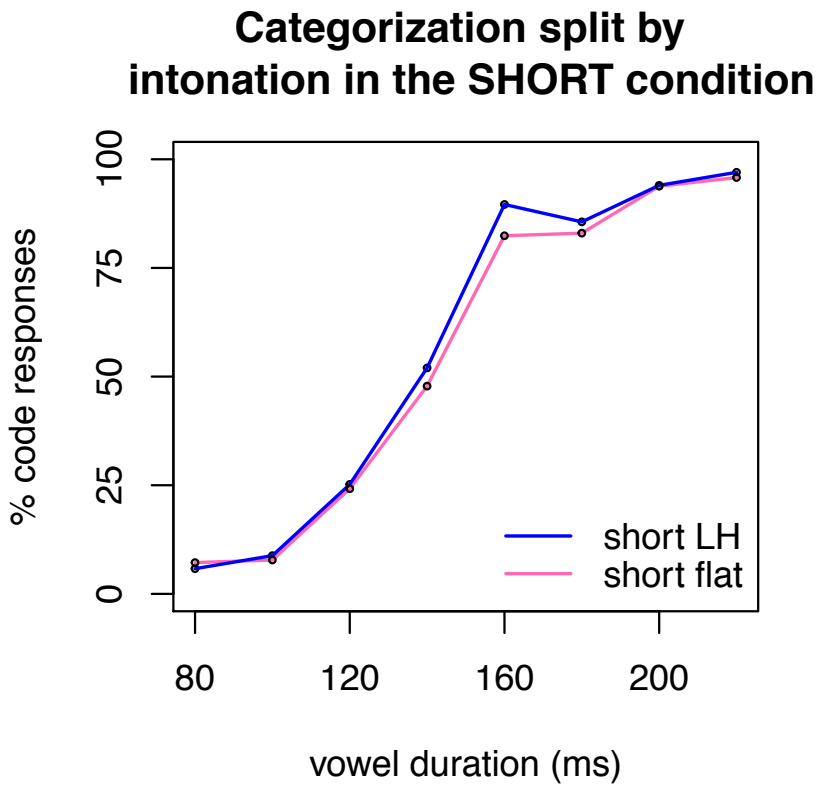

Figure 4: *Experiment 2 categorization split by intonation, in the SHORT condition.*

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 by 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 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.

4 General discussion

The results of the experiments reported here support the hypothesis that listeners interpret a compressed boundary tone (LH intonation in the SHORT condition) as an increase in speech rate, modulating subsequent categorization of VOT (Experiment 1) and vowel duration (Experiment 2). This suggests more broadly that listeners are sensitive to intonational structure in computing speech rate, and that these effects extend to the perception of durational (segmental) contrasts. The present study thus further develops the conclusions reached by Mitterer et al. (2016), 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.

As outlined in section 2.5, the results from the present study are not entirely supportive of the claim originally made by Kim & Cho (2013), who argued that listeners might modulate categorization to compensate for their expectations about initial strengthening. In specific terms, the results from Experiment 1 did not show that the boundary-cueing L-H% caused listeners to shift to higher VOT values (i.e., longer VOT for a /p/ response) for a post-boundary stop (with a boundary cued only by pitch). The results from Experiment 2 replicated the findings of Experiment

1, and are not explainable as listener compensation for initial strengthening, as outlined in section 3. This finding is in line with Mitterer et al. (2016), whose results showed that listeners did not adjust categorization when F0 information was removed, offering converging evidence that listeners are not sensitive to F0 modulations in compensation for initial strengthening.

However, the present results do not necessarily rule out the possibility that listeners do compensate for initial strengthening as originally argued by Kim & Cho. They simply show that intonational information is used by listeners in computing speech rate, not in triggering compensation for initial strengthening. Further investigation of 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 also make several concrete predictions for further study. Because intonational systems are language specific, the effect of tonal contours in cueing speech rate would be expected to vary across languages with different intonational systems and categories. One promising test case is Seoul Korean, where a falling-rising pitch pattern similar in shape to English L-H% can occur at the end of an accentual phrase (AP) without substantial lengthening (e.g. Jun 1993, 1995, 1998, 2005). Because the contour does not necessarily co-occur with lengthening, as it does in English, it would be expected that in analogous experiments with speakers of Seoul Korean and Korean stimuli an intonation-based shift in categorization would not be observed, because the

contour would not indicate an increased speech rate when distributed over a short vowel. This sort of cross-linguistic research is another important step in exploring how language-specific intonational systems affect the perception of fine-grained phonetic detail via rate normalization.

A related theoretical implication is that speech rate normalization, which can be seen as being early-stage, domain-general speech processing (as stated by Mitterer et al., 2016), is mediated to some degree by listeners' knowledge of intonation. This may seem contradictory. However, as mentioned in section 1.1, 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-Berke 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 category boundary shifting rate effects (of the kind presented here) have also shown that linguistic factors play a role. For example, Bosker & 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 the 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 need to 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 & 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 et al., 2016; Xu et al., 2006), and even the early-stage neural representation of pitch in the auditory brain stem (Krishnan et al., 2005, 2009a, 2009b). 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 understood (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 (L-H%) 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 language intonation 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.

Appendix A: Experiment 1

Table 3: *Output from the logistic regression in Experiment 1 (referred to as Model 1). Approximate p values are also shown in the right-most column.*

	B (SE)	z value	p value
(Intercept)	1.51(0.11)	13.18	< 0.001
VOT	2.77 (0.16)	17.42	< 0.001
intonation	-0.06 (0.03)	-2.24	0.03
Length	-0.24 (0.05)	-4.55	< 0.001
VOT:intonation	0.022 (0.05)	0.48	0.63
VOT:length	0.35 (0.05)	6.55	< 0.001
intonation:length	0.04 (0.03)	1.39	0.16
VOT:intonation:length	-0.08 (0.03)	-2.44	0.015

Table 4: *Variance and standard deviation for the random effect structure of Model 1.*

	variance	sd
(Intercept)	0.60	0.78
VOT	1.16	1.08
length	0.09	0.30
VOT:F0	0.04	0.19
VOT:length	0.04	0.21

Table 5: *Model 1, determination of random effect structure. The determination of the random effect structure of the model began by using the maximum possible random slopes with by-subject intercepts, and then successively simplifying the random effect structure until convergence was achieved. R notation for interactions is used. The slope with the lowest variance was removed successively, until the model converged. In the table below, the relevant models are given labels, and their random effect structure is summarized.*

Model	Converged?	Random slopes	Slope removed from previous model
Model A	N	VOT*intonation*length	---
Model B	N	VOT+ intonation +length+VOT: intonation +VOT:length	intonation:length
Model C	Y	VOT+length+VOT: intonation +VOT:length	intonation

Table 6: *Experiment 1 model comparisons. Tests of the simplified models against the maximally specified model and each other were carried out with the anova( ) function in R, which compares reduction in the residual sum of squares across models and computes whether this difference is statistically significant. This test determined the difference to be non-significant, meaning the simplifications used to reach convergence did not change the model significantly from the original, non-converging model.*

Comparison	χ^2	df	p value
Model A ~ Model B	11.13	15	0.74
Model B ~ Model C	5.58	6	0.47
Model A ~ Model C	16.714	21	0.73

Table 7: *Assessing the three way interaction with the lsmeans post hoc test. Below 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. Results show the asymmetry across length conditions. Significant p values are bolded. Note this asymmetry also holds independently of VOT, where lsmeans found a significant effect of intonation in the* SHORT *condition overall (Estimate = -0.187, z ratio = -2.53, p =0.011), and no effect in the* LONG *condition (Estimate = -0.04, z ratio = -0. 639, p = 0.52).*

VOT	Conditions compared	Estimate (SE)	z ratio	p value
0ms	SHORT LH – SHORT FLAT	-0.49(14)	-3.48	**0.003**
	LONG LH – LONG FLAT	0.13(0.17)	0.75	0.88
5ms	SHORT LH – SHORT FLAT	-0.42(0.11)	-3.84	**0.0007**
	LONG LH – LONG FLAT	0.08(0.13)	0.67	0.91
10ms	SHORT LH – SHORT FLAT	-0.36(0.08)	-4.26	**0.001**
	LONG LH – LONG FLAT	0.05(0.10)	0.51	0.95
15ms	SHORT LH – SHORT FLAT	-0.29(0.06)	-4.35	**0.001**
	LONG LH – LONG FLAT	0.01(0.07)	0.16	0.99
20ms	SHORT LH – SHORT FLAT	-0.22(0.07)	-3.32	**0.005**
	LONG LH – LONG FLAT	-0.03(0.06)	-0.40	0.98
25ms	SHORT LH – SHORT FLAT	-0.15(0.08)	-1.81	0.27
	LONG LH – LONG FLAT	-0.07(0.08)	-0.79	0.86
30ms	SHORT LH – SHORT FLAT	-0.08(0.11)	-0.76	0.87
	LONG LH – LONG FLAT	-0.10(0.11)	-0.93	0.79
35ms	SHORT LH – SHORT FLAT	-0.02(0.14)	-0.11	0.99
	LONG LH – LONG FLAT	-0.14(0.14)	-0.972	0.77
40ms	SHORT LH – SHORT FLAT	0.05(18)	0.30	0.99
	LONG LH – LONG FLAT	-0.18(0.18)	-0.98	0.76
45ms	SHORT LH – SHORT FLAT	0.12(0.21)	0.58	0.94
	LONG LH – LONG FLAT	-0.21(0.22)	-0.99	0.76

Appendix B: Experiment 2

Table 8: *Output from the logistic regression in Experiment 2 (referred to as Model 2).*

	B (SE)	z value	p value
(Intercept)	0.54 (0.08)	6.58	< 0.001
vdur	2.67 (0.13)	20.33	< 0.001
intonation	-0.04(0.02)	-1.69	0.09
length	-0.10 (0.03)	-2.78	0.005
vdur:intonation	0.04 (0.03)	-1.30	0.19
vdur:length	-0.01 (0.05)	-0.20	0.80
intonation:length	0.06 (0.02)	2.71	0.007
vdur:intonation:length	-0.04 (0.03)	1.35	0.18

Table 9: *Variance and standard deviation for the random effect structure of Model 2.*

	variance	sd
(Intercept)	0.29	0.54
length	0.03	0.17
vdur	0.76	0.87
length:vdur	0.02	0.14

Table 10: *Model 2, determination of random effect structure*

Model	Converged?	Random slopes	Slope removed from previous model
Model A	N	vdur* intonation *length	---
Model B	N	vdur + intonation + length + vdur:intonation + vdur:length + length:intonation	vdur:intonation :length
Model C	N	vdur + intonation + length + vdur:intonation + vdur:length	length:intonation
Model D	N	vdur + length + vdur:intonation + vdur:length	intonation
Model E	Y	vdur + length +vdur:length	vdur:intonation

Table 11: *Experiment 2 model comparisons*

Comparison	χ^2	df	p value
Model A ~ Model B	0.96	8	0.99
Model A ~ Model C	10.08	15	0.81
Model A ~ Model D	17.81	21	0.66
Model A ~ Model E	23.12	26	0.63
Model B ~ Model C	9.12	7	0.24
Model B ~ Model D	16.85	13	0.21
Model B ~ Model E	22.15	18	0.22
Model C ~ Model D	7.73	6	0.26
Model C ~ Model E	13.03	11	0.29
Model D ~ Model E	5.30	5	0.38

References

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*, 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., Maechler, M., Bolker B., & Walker, S. (2015). 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., & 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*.

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

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.

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.

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). Université Paris III Sorbonne Nouvelle, Paris.

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.

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.

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 disserataion). The Ohio State University.

Jun, S.-A. (1995). "Asymmetrical prosodic effects on the laryngeal gesture in Korean" in B. Connell and A. Arvaniti (eds.) *Phonology and phonetic evidence: Papers in Laboratory Phonology* IV: 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.

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. V. (2016). Least-Squares Means: The R Package lsmeans. Journal of Statistical Software, 69(1), 1-33.

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.

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 Commun.*, *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., & 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. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.

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., & Shattuck-Hufnagel, S. (2007). Multiple targets of phrase-final lengthening in American English words. *Journal of Phonetics*, *35*(4), 445–472.

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.

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.