Listening fast and slow: Exemplar matching or speech rate normalization in Japanese

Yoonjung Kang and Timothy Gadanidis (University of Toronto)

Introduction: Speech is highly variable: the same target structure can be realized very differently depending on the linguistic context and the speaker. Despite these variabilities, speech communication is largely successful because listeners take into account the effect of linguistic context and the speaker and calibrate their perception to arrive at the intended message (Hay et al. 2006a; Hay et al. 2006b; Johnson et al. 1999; Mitterer 2006; Niedzielski 1999; Schertz et al. 2019; Strand 1999; Yu 2010). One major source of speech variation is speech rate—variation due to how fast one speaks. The absolute length of segments differs depending on the speech rate, and this poses a particular challenge for sound contrasts that rely on duration. Scholars have found that speakers categorize target contrasts adjusting for the contextual speech rate (Kidd 1989; Miller 1987; Miller, Green, & Reeves, 1986; Mitterer 2018; Nagao and de Jong 2007; Newman and Sawusch 1996). However, evidence for compensation is not consistently found, depending on the proximity of speech rate cues to the target structure, the type of target structure, and across individuals (Heffner et al. 2017; Kang et al. 2018; Li and Kang 2018; Nakai and Scobbie 2016; Ting and Kang 2019). In Japanese, which has phonemic consonant and vowel length, it has been argued that all that is needed for the listener to distinguish long and short consonants is the preceding and following vowels (Amano & Hirata 2010; Hirata & Whiton 2005), while Hirata and Lambacher (2004) suggest that the vowel contrast is affected by the rate of distal speech material. This study examines how Japanese listeners incorporate different contextual cues for duration in calibrating their perception of consonant and vowel length.

Goals: The current study addresses the following questions through a series of experiments:

- Do listeners use speech rate normalization when perceiving vowel and stop quantity contrasts in Japanese?
- What are the differences (if any) in rate normalization for vowels vs. stops?
- Is rate normalization exemplar-based (i.e., stop perception is primarily modulated by duration of other stops in ambient speech) or dependent on global speech rate (i.e., stop perception is modulated by duration of all segments in ambient speech)?

Methods: We built 5 online experiments using jsPsych (de Leeuw 2015) and recruited participants via CrowdWorks, a Japanese crowdsourcing site. We created duration continua based on four vowel and four consonant real-word minimal pairs (e.g., k**a**do ‘corner’ vs. k**aa**do ‘card’; i**k**a ‘below’ vs. i**kk**a ‘family’) and spliced the target word to a carrier sentence: *Takeuchi-san wa totemo odayaka ni [word] to hatsuonshita* “Mr. Takeuchi very calmly pronounced [word].” The duration continua were created from productions of both a long and short member of each pair and were varied in the target segment length only with the rest of the word was held constant. The steps of each continuum were selected based on a 10-person pilot such that the odds of response are comparable across all continua. The carrier sentence was manipulated to be either 20% faster or 20% slower than the original production, by altering all segments equally, consonants only (Exp 2-4), or vowels only (Exp 5). Participants listened to the stimuli embedded in a carrier sentence and selected which word they heard. Mixed-effects logistic regression models were used for statistical analyses.

Results:
Exp 1 (2 segment **TYPES** [vowels vs. consonants] * 4 minimal word pairs * 2 original word **LENGTH** [long vs. short] * 9 **STEPS** *2 speech **RATES** = 288, 30 participants): We investigated whether and to what extent rate normalization occurs for both vowels and stops, by manipulating the rate of a carrier sentence and testing whether this biases participants' perceptions. We find that rate normalization does occur for both categories, but that the normalization effect is stronger for stops than for vowels (Figure 1), despite the fact that the local CV ratio cue remains invariable across speech rate condition contra Amano & Hirata (2010).

Exp 2 (2 segment **TYPES** [vowels vs. consonants] * 4 minimal word pairs * 2 original word **LENGTH** [long vs. short] * 5 **STEPS** *2 speech **RATES** * 2 rate **MANIPULATION** methods [all segs vs. vowels only] = 360, 40 participants): We probed whether listeners use overall speech rate or category-specific exemplar for duration perception adjustment by altering the rate of the carrier sentence by 20%, by either manipulating only vowels (leaving consonant duration the same) or by manipulating all segments equally. We find little difference between the two manipulation types: rate affects perception for both vowels and stops. However, post-hoc tests show that rate has a weaker effect for stops in the vowel-only manipulation condition, compared to the overall manipulation condition. The effect of rate on vowels does not vary significantly by manipulation condition (Figure 2)

Exp 3 & 4: One hypothesis is that pitch accent may play a larger role for vowels, making them less sensitive to speech-rate variation in general, since pitch is a stronger cue, so that the finer distinction between the two manipulation conditions is not affected. This hypothesis was tested in Exp 3, where the pitch cues were removed by flattening the pitch in all stimuli. We still found a difference between manipulation types for stops, and no difference between manipulation types for vowels. In Exp 4, we replicated Exp 3 using a blocked design for the rate modification, rather than a pseudorandom order to draw attention to the category-specific rate and, potentially, allow a category-specific effect for vowels to surface. The results are largely the same as Exp 3; whether the stimuli were presented in a blocked or unblocked order did not have a significant effect.

Exp 5: In Experiment 5, we invert the carrier-rate manipulation from Experiments 2–4 by increasing or decreasing the rate of the carrier by 20% by either manipulating consonants only (leaving vowel duration the same) or by manipulating all segments equally. Similar to Exp 2-4, we found that when only consonants are manipulated, the effect of rate for stops is greater than for vowels but vowel perception showed a consistent rate effect regardless of manipulation condition, despite the fact that in consonants only condition vowel duration in carrier sentence did not differ across rate conditions. This is the result that we would expect if rate normalization is category-specific for consonants but general for vowels, as suggested by results of previous experiments. (Figure 3)

Discussion: Our results indicate that rate normalization may be, at least in part, category specific. The effect of rate normalization is weaker for stops (but still exists) when only vowel durations are manipulated (Experiments 2–4); and it is stronger when stop durations are manipulated at a higher extent (Experiment 5). This suggests that one of the cues listeners attend to when perceiving the duration of a stop is the duration of other stops in that category

in the speech stream. However, because a weaker effect of rate still exists when stops are not modified at all, there still seems to be a general effect of speech rate, such as neural entrainment (e.g., Kösem et al. 2018), at play as well. This partial category-specificity does not arise for vowels: the effect of vowels does not vary depending on whether all segments are manipulated or only vowels are manipulated. For vowels, the overall speech rate appears to be more important, regardless of whether vowels, consonants, or both are manipulated.

Figure 1: Experiment 1

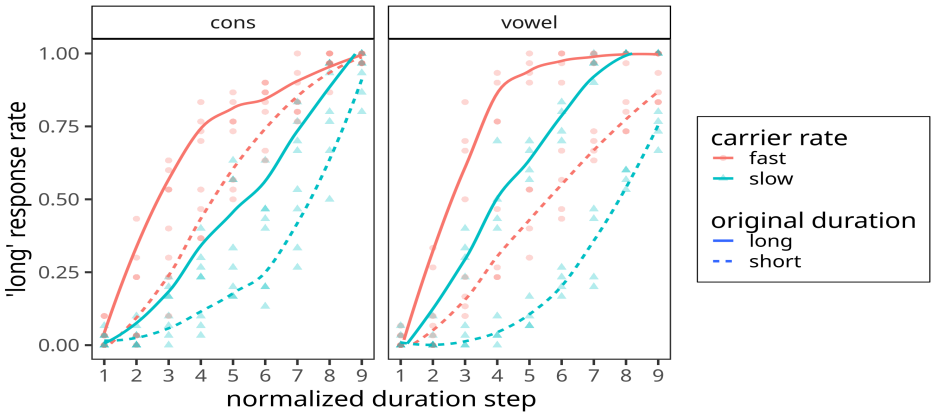

Figure 2: Experiment 2

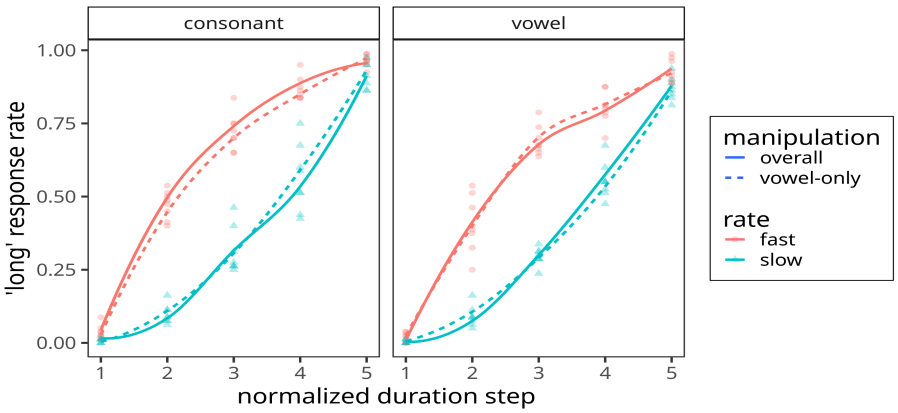

Figure 3: Experiment 5