Tracking speaker-specific speech rate: Habitual vs. local influences on English stop voicing*

Connie Ting[1,2] & Yoonjung Kang[3]

[1]McGill University, [2]Centre for Research on Brain, Language and Music, [3]University of Toronto Scarborough
connie.ting@mail.mcgill.ca, yoonjung.kang@utoronto.ca

ABSTRACT

Studies have shown listeners can track individual speakers' speech rate and adjust their perception of duration contrasts. However, studies show mixed results regarding whether listeners adjust their perception similarly for vowels vs. consonants. One study showed that after hearing a dialogue between a fast and slow speaker, listeners adjusted their perception of German vowel length contrasts (/a/ vs. /a:/) in a speaker-specific way. A replication study showed no such rate effect for the English voicing contrast. In our study, English listeners heard a dialogue between a fast and slow speaker, with a greater rate difference compared to previous studies, followed by two identification tasks. The first included isolated stop-initial syllables (e.g. pig~big) manipulated along a VOT continuum. The second included the same syllables embedded in a fast or slow carrier sentence. Our results showed listeners adjusted their perception of word-initial stops in both the rate-manipulated carrier sentences and dialogue condition.

Keywords: speech rate, VOT, speech perception, individual variation

1. INTRODUCTION

Research has shown that speakers can vary substantially in speech rate not only across speakers but also within individuals [1, 2]. Speech rate variation is particularly pertinent to durational properties, since faster speech compresses utterance duration while slower speech expands duration. Considering the amount of variation that occurs, listeners must be able to normalize for speech rate to correctly interpret the speech signal from their interlocuter. That is, listeners need to perceive durational acoustic cues relative to the surrounding speech rate and/or to the speech rate that the listener associates with a particular speaker.

The effects of speech rate on listeners' perception have been studied at least as far back as the 1960's [3]. In general, the reported effect is that a slower contextual speech rate leads listeners to perceive an ambiguous unit of speech as relatively short, whereas a faster contextual speech rate leads listeners to perceive the same stimulus as relatively long [4, 5, 6, 7]. Previous work has also shown that listeners can track durational properties in a speaker-specific fashion [8].

Importantly, however, speech rate effects may differ across speech contexts. Many studies have shown that speech perception is influenced by relatively short adjacent contexts, such as an adjacent phoneme [7, 9] or a carrier sentence containing the target sound/word [10]. More recent studies have shown that listeners' perception can be influenced by the global or habitual speech rate of a speaker [11, 12]. That is, listeners can track the average speech rate of a speaker over a longer period of time. For example, Reinisch [12] examined speech rate effects in the context of a dialogue between two speakers. After hearing a 2-minute dialogue between two female native speakers of German, varying in rate (fast vs. slow) and order (first vs. second speaker), listeners completed a phonetic categorization task in which they categorized words of minimal pair continua differing in the /a/-/a:/ duration contrast. The stimuli were presented as words in isolation, intermixed across speakers. Their results showed that listeners' perception of the vowel contrast shifted depending on the speech rate of the speaker. Stimuli produced by the fast speaker in the preceding dialogue elicited more /a:/ responses compared to the slow speaker.

These results suggest that listeners are able to track speaker-specific speech rate information in a dialogue context and make use of this information in later perception. However, it is possible that the results are due to the presence of target vowels in the dialogue which could not be avoided, rather than perceptual adjustments due to speech rate. A similar study [13] aimed to replicate Reinisch's findings with English listeners' perception of a consonantal contrast, namely voice onset time (VOT). They presented listeners with a dialogue between two speakers, matching the rate manipulation of Reinisch's study [12], but crucially did not include any instances of stressed syllable-initial voiceless stops in the dialogue. Results from their categorization task showed no effect of speech rate.

Since these studies examined two different types of contrasts (vowels vs. consonants) and also differed in the presence or absence of the target structure in

the dialogue, there are multiple reasons why the two studies diverged in their findings. Given the lack of speech rate effect on a consonantal contrast, it remains unclear whether a speaker-specific speech rate effect can be found for a vowel length contrast using a dialogue which omits all instances of the target stimuli. In other words, it is possible that no speech rate effect emerges when listeners are not exposed to target stimuli in the dialogue, regardless of the kind of durational contrast being examined.

On the other hand, it is possible that the different results are due to the particular type of contrast being examined. One possibility is that the degree of rate difference contributes to the presence of rate normalization and that a greater rate difference is required to elicit rate normalization of consonantal contrasts than of vowel contrasts. That is, the lack of the speech rate effect found for the consonantal contrast may be attributed to a rate difference between speakers that was not salient enough.

Another difference between [12] and [13] is the former used multiple minimal pairs as target words while the latter used a single monosyllable word pair, making the task far less natural and more prone to decay of dialogue rate effects through the experiment.

The current study replicates [13] with two crucial modifications – making the rate difference larger and using multiple word pairs. In addition, we included another identification task following the main task, whereby the same word stimuli used in the main task is presented embedded in a rate-manipulated carrier sentence, where previous studies [14] have shown robust rate effects and we should expect the same. Significant rate effects in the sentential condition could assure us that the target stimuli manipulation is natural and the online participants are performing the tasks as intended. The results of our main tasks can thus be evaluated in a firmer footing.

2. METHODS

2.1. Participants

80 listeners were recruited through the Amazon Mechanical Turk (Mturk) platform and were paid for their participation. All listeners were self-identified native speakers of American English and reported normal speech and hearing.

2.2. Stimuli

Two male speakers (M1, M2) individually recorded a 460-word dialogue between two people, which was scripted such that no stressed syllable-initial voiceless stops were included. Each speaker recorded both roles of the dialogue (A and B) and were instructed to read the dialogue at a comfortable rate.

The dialogue recordings were segmented at phrase boundaries and labelled according to the speaker role (A or B). Phrase durations were measured to determine the natural speech rate for each speaker. Phrase durations were then manipulated to create two speech rate conditions (fast and slow), by compressing or expanding the speech to be 20% shorter or longer than the average of the two speakers' natural speech rate (exaggerated from rate conditions used in previous studies, which ranged from 10% shorter to 15% longer [12, 13, 14]). Manipulated phrases were spliced back together leaving 250 ms of silence between the utterances. After duration manipulation, the resulting dialogue was ~2 minutes. Four versions of the dialogue were used in this study such that each speaker was heard in each role (A and B) and at each speech rate (fast and slow).

Each speaker also recorded 5 repetitions of 8 words (4 pairs: time-dime, pig-big, toe-doe, pan-ban) in a carrier sentence: 'Now I will say ___'. A carrier sentence with the target word in final position was chosen following previous studies which have shown that VOT is perceived by listeners relative to preceding speech rate context [13, 14]. Measurements were made for the duration of the carrier sentence, as well as the closure, aspiration, vowel, and coda (when applicable) of the target word to determine the average duration of each speaker's natural production of the carrier sentence and target words. To make the sentence stimuli as comparable as possible, one repetition of the carrier sentence was extracted from each speaker's production such that the sentences sounded most similar between speakers in terms of overall pitch contour and included minimal pause between words. Base tokens of the chosen carrier sentences were created by manipulating the duration of the sentence as well as the closure duration between the carrier sentence and target word to be the mean duration across speakers.

For the target word stimuli, a base token for each word pair was created by splicing together the aspiration of a voiceless token and the vowel+coda from its voiced counterpart, all originally produced in the same carrier sentence context. The durations of the vowel and the coda were manipulated to be equal to the word-pair means across speakers. The VOT duration of base tokens was manipulated to create VOT continua ranging from 0 to 70 ms in 8 steps.

For sentential condition stimuli, the carrier sentence was manipulated to create two rate conditions (fast and slow) similar to the dialogue manipulation (20% faster or slower) and the target words were spliced back onto one of the rate-manipulated carrier sentences. In total, the manipulations were used to create 4 dialogue conditions, 64 words in isolation (8 VOT steps * 4

word pairs * 2 speakers) and 128 sentences (8 VOT steps * 4 word pairs * 2 speakers * 2 sentence rates).

Each speaker also recorded 5 repetitions of the 8 words in isolation. The third repetition for each word was used for control stimuli (16 total: 8 words * 2 speakers). These tokens were not manipulated and were used to represent naturally produced tokens.

A pretest was run to determine the range of the VOT continuum that would be sufficient in obtaining a balance of voiced vs. voiceless responses. 30 participants who did not take part in the main experiment participated online through Mturk and were paid for their participation. The final VOT continuum used for the main experiment ranged from 10 to 70 ms in 7 steps, as stimuli beyond these boundaries consistently elicited expected responses.

2.3. Procedure

The experiment was built and hosted on the Gorilla platform (gorilla.sc) [15]. Before beginning the experiment, listeners completed an audio test to ensure that the sound was playing properly and at a comfortable volume.

Each participant heard one of four versions of the dialogue. Each version was heard by 20 participants. After hearing the dialogue, participants completed two categorization tasks. The first presented words in isolation. On each trial, participants saw a fixation cross for 500 ms, with 100 ms of pause before and after the fixation cross. They then heard an individual word and were asked to click the corresponding word on the screen. The stimuli for the two speakers were presented intermixed. For each trial, two words were displayed on the screen: the button on the left side of the screen always corresponded with the word with a voiceless stop (e.g., time) while the button on the right side of the screen always corresponded with the voiced counterpart (e.g., dime). After each response by button click, the next trial would begin. Each stimulus was repeated 2 times, resulting in a total of 112 trials for each listener (7 VOT steps * 4 word pairs * 2 speakers * 2 repetitions).

After hearing the words in isolation, listeners were presented with a second categorization task with the same procedure, but with the target word embedded in a carrier sentence. Listeners saw a fixation cross, then heard the sentence 'Now I will say ___', and were asked to indicate the last word of the sentence by clicking the corresponding word on the screen. The buttons were arranged in the same way as the previous task. Each stimulus was repeated once, resulting in a total of 112 sentential trials (7 VOT steps * 4 word pairs * 2 speakers * 2 sentence rates).

Note that speech rate was a between-subject factor in the dialogue condition; the same speaker was heard as fast by some participants but as slow by others. In the sentential condition, all participants heard the same stimuli, and rate was a within-subject factor.

Finally, listeners were presented with a short block of control stimuli with the same procedure as the previous tasks. The control block consisted of 16 tokens (8 words * 2 speakers) randomized. These were unmanipulated natural tokens, intended to elicit the correct response by listeners and used as a control such that listeners who did not correctly identify all of the natural tokens were excluded from the analysis (n=4). The full experiment took approximately 20 minutes to complete.

3. RESULTS

All statistical analyses were conducted in R [16]. Starting with the dialogue condition, in which listeners heard the dialogue then categorized words in isolation, Figure 1 shows the proportion of voiceless responses over the VOT continuum for the fast versus slow speaker in the dialogue aggregated over all four versions of the dialogue. Figure 1 shows a shift in category boundary, particularly in the ambiguous region of the VOT continuum (~30 ms), such that the same VOT duration was perceived more often as voiceless for speakers' fast speech than slow speech. This is represented by the solid line, representing fast speech, appearing above the dotted line, representing slow speech.

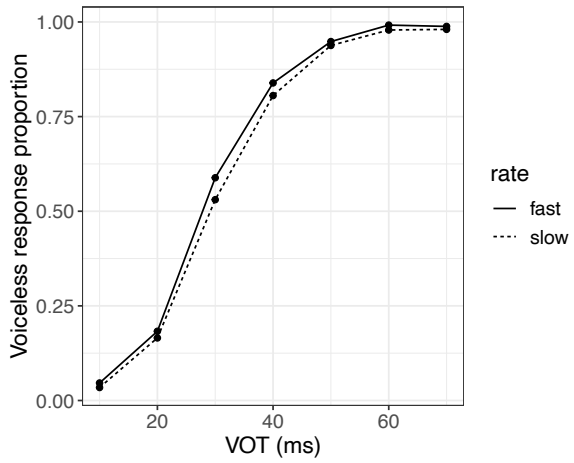

Figure 1: Proportion of voiceless responses over the VOT continuum for the fast (solid line) versus slow (dotted line) speaker in the dialogue.

A logistic mixed-effects model was fit using the *glmer* function of the *lme4* package [17]. The model included response (voiceless coded as 1, voiced as 0) as a dependent variable and VOT (ms, centred), SPEAKER (M1 = -0.5, M2 = 0.5), SPEECH RATE (fast = -0.5, slow = 0.5), and interactions between SPEECH

RATE and VOT and between SPEECH RATE and SPEAKER, as fixed factors. By-PARTICIPANT and by-WORD PAIR random intercepts were included, in addition to by-PARTICIPANT random slope adjustments to VOT. The results of the model showed a significant effect of VOT (b_{VOT} = 0.20, z = 27.41, p <0.001), with more /p/ responses as VOT duration increased, as expected. Results also showed a significant effect of SPEAKER ($b_{Speaker}$ = -0.80, z = -9.94, p = <0.001), indicating that speaker M2 had significantly less voiceless responses than speaker M1. The results also showed a significant effect of SPEECH RATE ($b_{Speech\ Rate}$ = -0.40, z = -4.02, p = <0.001), indicating that when compared across dialogue conditions, a speaker's fast speech condition elicited more /p/ responses compared to their slower condition. There were no significant interactions ($b_{Speech\ Rate * VOT}$ = -0.01, z = -1.65, p = 0.099; $b_{Speech\ Rate * Speaker}$ = 0.06, z = 0.15, p = 0.885).

We turn next to the sentential condition, in which listeners heard target words embedded in a carrier sentence which varied in rate (fast or slow). Figure 2 shows the proportion of voiceless responses over the VOT continuum for the fast versus slow sentential contexts. The results of this task as plotted in Figure 2 show a consistent difference in VOT categorization due to speech rate, such that a faster sentential context elicited more voiceless responses. This is shown by the solid line, representing fast speech, appearing above the dotted line, representing slow speech.

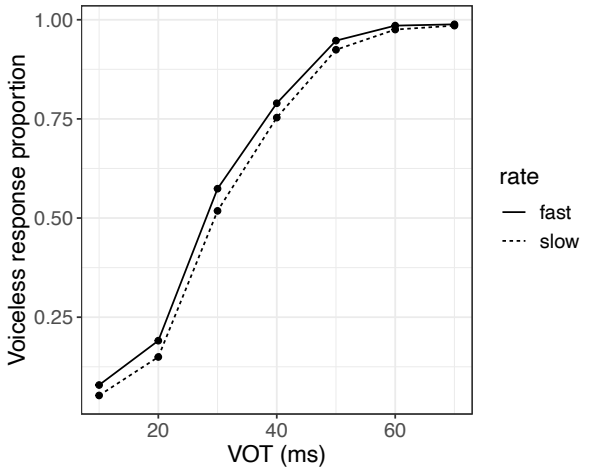

Figure 2: Proportion of voiceless responses over the VOT continuum for fast (solid line) versus slow (dotted line) sentential contexts.

The model specification for the sentential task was identical to the dialogue model. The results showed a significant effect of VOT (b_{VOT} = 0.17, z = 27.60, p <0.001), with more /p/ responses as VOT duration increased, as expected. Results also showed a significant effect of SPEAKER ($b_{Speaker}$ = -0.25, z = -3.35, p = <0.001), indicating that speaker M2 had significantly less voiceless responses than speaker M1. The results also showed a significant effect of SPEECH RATE ($b_{Speech\ Rate}$ = -0.33, z = -3.72, p <0.001), indicating that the fast speech rate sentences elicited more voiceless responses compared to the slow speech rate sentences. There were no significant interactions ($b_{Speech\ Rate * VOT}$ = 0.0003, z = 0.05, p = 0.962; $b_{Speech\ Rate * Speaker}$ = -0.24, z = -1.61, p = 0.108).

4. DISCUSSION

The current study tested whether listeners keep track of individual speakers' speech rate and whether rate normalization differs across speech rate contexts. In the first part of the experiment, listeners heard a dialogue between two male speakers and then categorized individual words. The dialogue was created such that it provided listeners with each speaker's speech rate in direct contrast (one fast, one slow) and did not include any of the target stimuli (stressed word-initial stops). Results from the categorization task following the dialogue showed a significant effect of speech rate, suggesting that the speaker-specific speech rate affected listeners' perception of the English voicing contrast.

In the second part of the experiment, listeners categorized target words which were embedded in a carrier sentence that was either fast or slow. The results of this task also revealed a significant effect of speech rate, showing that faster speech in the sentential context elicited more voiceless responses. The significant effect of speech rate found in both contexts suggests that rate normalization in listeners' perception of consonantal contrasts can occur in both local (sentential) and habitual (dialogue) contexts.

Note that the rate manipulation in this study was greater than that used in previous studies (see [12] for vowels, [13] for consonants) allowing us to test whether a more exaggerated rate difference would lead to a clear speech rate effect. Given that a smaller speech rate difference has led to rate effects for vowel contrasts but not consonant contrasts [12, 13], the significant rate effect in the current study for English VOT using a more exaggerated rate difference could indicate that speech rate normalization does not behave in exactly the same way across different durational contrasts. Additionally, the significant rate effects found for vowel contrasts might be attributable to the presence of target structure in [12], rather than speech rate alone. More work is required to test the presence and degree of rate normalization across speech rate contexts and different durational contrasts in the absence of target structure tokens in the rate-defining speech context.

5. REFERENCES

[1] Miller, J. L., Grosjean, F., Lomanto, C. 1984. Articulation rate and its variability in spontaneous speech: A reanalysis and some implications. *Phonetica* 41, 215-225.

[2] Jacewicz, E., Fox, R. A., Wei, L. 2010. Between-speaker and within-speaker variation in speech temp of American English. *J. Acoust. Soc. Am.* 128, 839-850.

[3] Pickett, J., Decker, L. R. 1960. Time factors in perception of a double consonant. *Language and Speech* 3, 11-17.

[4] Miller, J. L. 1987. Rate-dependent processing in speech perception. In Ellis, A. W. (ed.), *Progress in the psychology of language*. London: Erlbaum, 119-157.

[5] Miller, J. L., Dexter, E. R. 1988. Effects of speaking rate and lexical status on phonetic perception. *J. Exp. Psychol. Hum. Percept. Perform.* 14, 369-378.

[6] Kidd, G. R. 1989. Articulatory-rate context effects in phoneme identification. *J. Exp. Psychol. Hum. Percept. Perform.* 15, 736-748.

[7] Sawusch, J. R., Newman, R. S. 2000. Perceptual normalization for speaking rate: II. Effects of signal discontinuities. *Perception & Psychophysics* 62, 285-300.

[8] Allen, J. S., Miller, J. L., DeSteno, D. 2003. Individual talker differences in voice-onset-time. *J. Acoust. Soc. Am.* 113, 544-552.

[9] Summerfield, Q. 1981. Articulatory rate and perceptual constancy in phonetic perception. *J. Exp. Psychol. Hum. Percept. Perform.* 7, 1074-1095.

[10] Dilley, L. C., Pitt, M. A. 2010. Altering context speech rate can cause words to appear or disappear. *Psychological Science* 21, 1664-1670.

[11] Baese-Berk, M. M., Heffner, C. C., Dilley, C. L., Pitt, M. A., Morrill, T. H., McAuley, J. D. 2014. Long-term temporal tracking of speech rate affects spoken-word recognition. *Psychological Science* 25, 1546-1553.

[12] Reinisch, E. 2016. Speaker-specific processing and local context information: The case of speaking rate. *Applied Psycholinguistics* 37, 1397-1415.

[13] Ting, C., Kang, Y. 2019. The effect of habitual speech rate on speaker-specific processing in English stop voicing perception. *Proc. 19th ICPhS* Melbourne, 3230-3234.

[14] Toscano, J. C., McMurray, B. 2015. The time-course of speaking rate compensation: effects of sentential rate and vowel length on voicing judgments. *Language, Cognition and Neuroscience* 30(5), 529-543.

[15] Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., Evershed, J. K. 2019. Gorilla in our midst: An online behavioral experiment builder. Behavior Research Methods.

[16] R Core Team. 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

[17] Bates, D., Mächler, M., Bolker, B., Walker, S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67(1), 1–48.

[18] Fox, J., Weisberg, S. 2019. An R Companion to Applied Regression, Third edition. Sage, Thousand Oaks CA.

[19] Newman, R. S., Sawusch, J. R. 2009. Perceptual normalization for speaking rate: III. Effects of the rate of one voice on perception of another. *J. Phon.* 37, 46-65.

*This study was funded by the Social Sciences and Humanities Research Council of Canada (#435-2020-0209).