THE EFFECT OF HABITUAL SPEECH RATE ON SPEAKER-SPECIFIC PROCESSING IN ENGLISH STOP VOICING PERCEPTION

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

[1]University of Toronto; [2]University of Toronto Scarborough
connie.ting@mail.utoronto.ca; yoonjung.kang@utoronto.ca

ABSTRACT

This study investigates listeners' ability to track individual talkers' habitual speech rate in the context of a short dialogue and adjust their perception of durational contrasts. Previous studies found that such adjustment is possible for German vowel length contrasts (/a/ vs. /a:/). The results, however, allow for alternative interpretations suggesting that the adjustment is due to the tokens of (target) vowels heard in the dialogue rather than a global calibration of the perception based on talker's habitual speech rates or that the need for rate normalization itself differs by language and contrast. In our study, English listeners were presented with a dialogue between a fast and a slow talker, containing no stressed syllable-initial voiceless stops, followed by an identification task in which listeners categorized /pi/-/bi/ syllables manipulated to differ along a VOT continuum. Our results found that when the dialogue did not include any instances of the target structure, listeners' response did not differ systematically depending on the talker's habitual speech rate.

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

1. INTRODUCTION

In everyday communication, listeners deal with highly variable speech. The same words can result in a range of differences based on a variety of factors such as linguistic environment, social context, and the speaker, among others. One source of variation that creates differences in the realization of durational acoustic cues, is changes in speaking rate. Languages make use of durational differences to contrast between words. In English, for example, voice onset time (VOT), the interval between the release of a stop and onset of vocal cord vibration, is a durational contrast that differentiates voiced from voiceless stop phonemes. Speech rate effects have been found for stop voicing contrasts such that as speech rate slows, VOT for voiceless stops increases systematically [7, 13, 14, 26]. This poses a potential problem for listeners because a slow speaker's realization of /b/ may have similar VOT values as a fast speaker's realization of /p/.

Research has shown that listeners are sensitive to contextual variations such as speech rate and are able to compensate for this variation by tuning their perception of VOT relative to the given speech rate [9, 11, 12, 13, 22]. Since VOT increases as speech rate slows, the same VOT was more often perceived as /b/ in slow speech, but as /p/ in fast speech. Similar speech rate effects have been found in the perception of vowel duration [20, 21].

Researchers have also examined speech rate effects over more global contexts. More specifically, studies have investigated whether listeners are able to track the speech rate of a speaker over an extended period of time, also referred to as habitual or global rate, rather than the speech rate of a carrier sentence directly preceding a target word, also referred to as local rate [2, 10]. Results demonstrated that listeners tracked variation in the overall speech rate of an individual speaker over an extended period of time and that their knowledge of the speakers' habitual speech rate influenced their speech perception.

The focus of this research has largely been on situations in which there is only one speaker. That is, listeners appear to make use of speech rate information when no other source of speech rate information is present. However, in daily life, listeners are often faced with situations in which multiple people are speaking. Since each speaker provides unique speech rate information, the process of rate normalization relies on the listeners' ability to track individual speakers separately and make use of their knowledge of speaker-specific properties in perception. Previous research suggests that listeners are able to track duration properties in a speaker-specific fashion. Namely, listeners have been shown to remember whether a certain speaker has a tendency to produce /p/ with a short VOT whereas a different speaker produces /p/ with a long VOT [1]. Such evidence suggests that it is likely the case that listeners also track speaker-specific rate information to facilitate speech perception.

Reinisch [21] sought to extend these findings by testing speaker-specific effects of speech rate on listeners' vowel length perception in German. The study examined speech rate effects in the context of a conversation between two speakers. In the first of two experiments, listeners heard a 2-minute dialogue between two female native speakers of German,

varying in rate (fast vs. slow) and order (first vs. second speaker). Following the dialogue, listeners completed a phonetic categorization task in which they categorized words of minimal pair continua differing in the /a/-/a:/ duration contrast and were asked to indicate which word they heard. Results of the experiment showed that listeners were able to retain speech rate information, resulting in a shift in perception of the vowel contrast depending on the speech rate of each individual speaker. As expected, more /a:/ responses were found for the fast speaker than the slower speaker.

In the second experiment, different listeners heard the same dialogue, and similarly performed a categorization task. However, the target words in the second experiment were not presented in isolation but were instead presented in a carrier sentence that was produced in either a fast or slow speech rate. Thus, in addition to the speakers' habitual rate experienced in the dialogue, the carrier sentence provided listeners with local rate information. Results of the second experiment showed no significant effect of habitual rate when local rate information was provided, and an overall stronger effect of local rate than habitual rate was found.

These results are restricted to the context of vowel duration contrasts. However, it is not the case that all durational contrasts are affected by speech rate in the same way and for all speakers. For example, it has been shown that change in duration due to speech rate is reflected primarily in changes in vowel duration rather than consonant duration [6, 18]. Moreover, VOT values associated with voiced stops are less affected by speech rate compared to those of voiceless stops [7, 14]. Furthermore, there exists individual variation such that while some speakers are sensitive to rate change in their perception and production of VOT, others are not [1, 7, 24].

Thus, to improve our understanding of how listeners deal with speech rate variation in speech perception, it is important to also investigate such effects in consonantal contrasts, such as VOT. More specifically, it remains to be tested whether habitual speech rate effects can be found in the perception of consonantal contrasts when listeners are tasked with attending to two speakers in a dialogue.

In addition to extending previous results, it is crucial to understand how such an effect of habitual speech rate may be different between vowel and consonantal contrasts. Vowels are ubiquitous in speech, such that any amount of exposure to a speaker's speech rate provides ample tokens of vowels to be used for comparison in later perception. For this reason, it remains ambiguous whether the effect of habitual rate found by Reinisch [21] is indeed evidence for general information about speech rate or whether it's the result of durational properties of the vowel tokens contained in the exposure. Thus, the current study aims to replicate the study by Reinisch [21] using consonantal contrasts, namely, VOT, to provide a more stringent test of speakers' habitual rate effect on subsequent speech perception.

2. METHODS

2.1. Participants

A total of 115 listeners participated in this study. All speakers were self-identified native speakers of American English and reported normal speech, hearing, and vision. Speakers were recruited online using Amazon Mechanical Turk (MTurk) and were paid for their participation.

2.2. Stimuli

A 324-word dialogue between two speakers was scripted such that no stressed syllable-initial voiceless stops were included. Two male speakers (M1 and M2) recorded both roles of the dialogue 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-turn (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), such that the fast version was compressed to be 15% shorter, and the slow version was expanded to be 10% longer than the average of the two speakers' natural speech rate. Manipulated phrases were spliced back together leaving 250 ms of silence between utterances. The amount of rate change and inter-utterance gap were chosen to reflect that of the similar study by Reinisch [21], while ensuring that the resulting dialogue was both natural and distinct enough to be recognized as fast and slow. After durational manipulation, the resulting dialogue was 1 minute and 56 seconds. Four versions of the dialogue were created such that each speaker was heard in each role (A and B) and speech rate (fast and slow).

Each speaker also recorded 10 repetitions of the words 'bee' /bi/ and 'pee' /pi/. Speakers were asked to repeat each word 10 times with sufficient pause in between each utterance. VOT and vowel durations were segmented and measured to determine the average VOT and vowel duration for each speaker's natural production of /bi/ and /pi/. Base tokens for creating the stimuli were chosen to represent the speakers' natural production of /bi/ and pi/ as much as possible, based on each speaker's average vowel duration across their /bi/ and /pi/ production, and the speaker's average VOT duration across their /pi/

production. For each speaker, stimuli were made by splicing the aspiration of the speaker's /pi/ token onto the vowel of the speaker's /bi/ token. Using the concatenated utterance for each speaker, the VOT duration was manipulated to create a VOT continuum ranging from 0-70 ms in 15 steps. The vowel duration was kept constant at 400 ms, the average vowel duration between the two speakers.

A pretest was run to determine the range of the VOT continuum that would be sufficient in obtaining a balance of /pi/ and /bi/ responses. Ten participants who did not take part in the main experiment participated through MTurk's online system and were paid for their participation. The listeners' task was to listen to the manipulated stimuli separated in two blocks, one for each speaker, and indicate whether they heard a /pi/ or /bi/ syllable by clicking the corresponding button on the screen. The final VOT continuum used for the main experiment ranged from 0-50 ms in 11 equal steps, as stimuli over this range elicited mostly /pi/ responses during the pretest. All stimuli were manipulated using Praat's PSOLA algorithm [5].

2.3. Procedure

Participants completed the experiment through the MTurk online system. At the start of the experiment, participants were instructed to have a pair of headphones ready for the listening task. They were required to specify the model of the headphones before the online interface allowed them to continue with the experiment. Each participant heard one of four versions of the dialogue (2 roles * 2 rates). Participants were instructed to listen carefully to the dialogue because they would be asked to answer questions about what they heard afterwards. Once the dialogue was finished, participants completed the categorization task in which they were asked to listen to the speakers say either 'pee' or 'bee' and click the corresponding word on the screen to indicate which word they heard. The button on the left side of the screen was always 'pee' and the button on the right side of the screen was always 'bee'. The task was self-paced with no time line on response. After each response by button click, the next stimulus would play after 500 ms. The two speakers' word items were presented intermixed. Each stimulus was repeated 3 times, resulting in a total of 66 trials for each listener (11 VOT steps * 2 speakers * 3 repetitions). Stimuli were randomized for each participant with the restriction that no identical token be presented twice in a row.

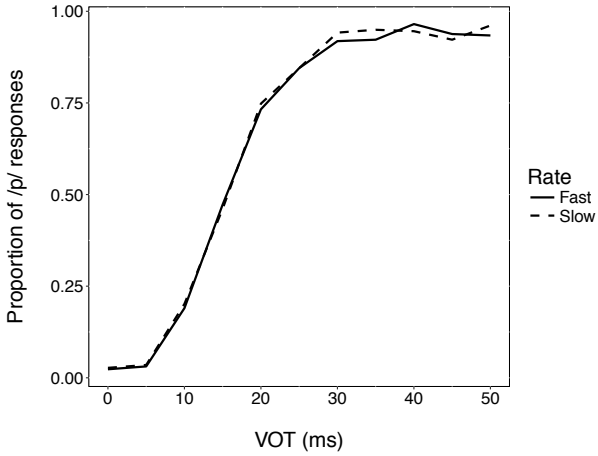
After the categorization task, participants were asked to answer a multiple-choice question asking where one of the speakers of the dialogue would be visiting. This information appeared within the last few sentences of the dialogue, and so was used as a method to exclude participants who were not paying attention during the experiment. Those who did not answer the question correctly were excluded from the analysis (n=20) and data from the remaining 95 participants were analyzed. The experiment took approximately 5 minutes to complete.

If listeners are able to keep track of individual speakers' habitual rate, we expect listeners to give more /p/ responses for the fast speech rate condition, compared to the slow speech rate condition. That is, if a speaker has a fast speech rate, a given VOT value will seem long relative to the speakers' habitual rate, and therefore elicit more /p/ responses. On the other hand, if a speaker has a slow speech rate, the same VOT value will seem short relative to the speakers' habitual rate, and therefore elicit less /p/ responses.

3. RESULTS

Participants whose perception did not show a significant effect of VOT, or who showed a significant effect of VOT in the opposite from expected direction (*less* /pi/ responses for longer VOT), were also excluded from further analysis (n=9). The final analysis included data from 86 listeners, with 20-23 listeners in each of the four dialogue conditions. Figure 1 shows the proportion of /p/ responses over the VOT continuum for the fast versus slow speaker in the dialogue. As shown in Figure 1, there is no clear difference between the responses given for the fast speaker (solid line) and the slow speaker (dotted line), suggesting no effect of speech rate across the two speakers.

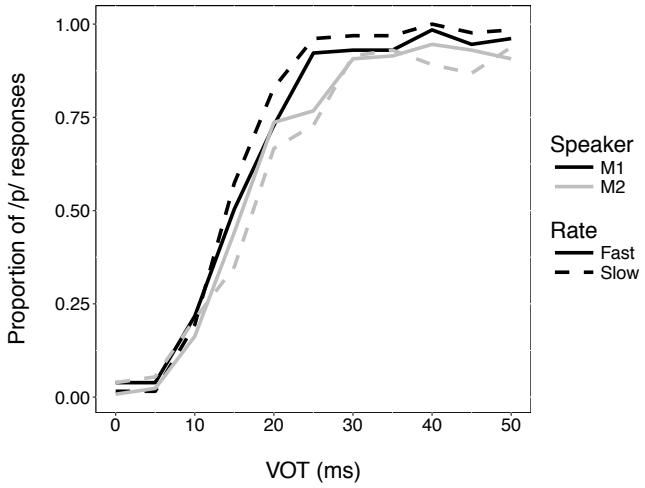
Figure 1: Proportion of /p/ responses over the VOT continuum for the fast (sold line) versus slow (dashed line) speaker in the dialogue.



When broken down by the individual speakers, more noticeable differences are observed. Figure 2 shows the proportion of /p/ responses over the VOT

continuum for the fast (solid line) versus slow (dotted line) speech rate, separately for the individual speakers (speaker M1 = black lines, speaker M2 = grey lines). First, there appears to be an overall difference between speaker M1 and speaker M2, such that there are more /p/ responses for speaker M1 than for speaker M2. This is indicated in Figure 2 with the black lines appearing above the grey lines, representing speaker M1 and speaker M2, respectively. Furthermore, the direction of the speech rate effect is in the expected direction for speaker M2, but in the opposite direction for speaker M1. As shown in Figure 2, the solid grey line, representing the fast speech rate condition for speaker M2, is slightly above the dotted grey line representing the slow speech rate condition for the same speaker. On the other hand, the solid black line, representing the fast speech rate condition for speaker M1, is slightly below the dotted black line representing the fast speech rate condition for that speaker.

Figure 2: Proportion of /p/ responses over the VOT continuum for fast (solid line) versus slow (dotted line) speech, separately for Speaker M1 (shown in black) and Speaker M2 (shown in grey).



Statistical analyses were conducted in R [19] and the *glmer* function of the *lme4* package [3] was used. A logistic mixed-effects model was fit with response (/p/ coded as 1, /b/ coded 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 their interaction, as fixed factors. The maximal random effect structure warranted by likelihood ratio tests was selected, which included by-PARTICIPANT random intercepts and by-PARTICIPANT random slope adjustments to VOT and SPEAKER. The results showed a significant effect of VOT (b_{VOT} = 0.33, z = 15.78, p <0.001), with more /p/ responses as VOT duration increased, as expected. Results also showed a significant effect of SPEAKER ($b_{Speaker}$ = -0.82, z = -3.99, p <0.001), indicating that speaker M2 had significantly less /p/ responses than speaker M1. However, there was no significant effect of SPEECH RATE ($b_{Speech\ Rate}$ = 0.07, z = 0.20, p = 0.740). The interaction of SPEAKER and SPEECH RATE was also not significant ($b_{Speaker * Speech\ Rate}$ = 0.04, z = 0.07, p = 0.947).

4. DISCUSSION

The present study sought to test whether and how listeners keep track of individual speakers' habitual rate in a short dialogue and make use of the information in a subsequent speech perception task. The dialogue between two male speakers provided listeners with each speaker's habitual speech rate in direct contrast. Results from the categorization task showed no effect of habitual speech rate, which suggests that either listeners did not track individual speakers' habitual rate or more likely, they did track individual's speech rate, but the information did not affect the perception of the English /p/-/b/ contrast. These results are inconsistent with previous findings and warrant further discussion.

In the present study examining the /p/-/b/ stop voicing contrast, listeners heard /pi/-/bi/ syllables in which the stimuli varied in VOT duration. Crucially, however, these stimuli contain a vowel, which was kept constant in duration across the two speakers. Note that the duration of the vowel following the target stop provides listeners with local rate information [22] or itself serves as a secondary cue for the stop voicing contrast [24]. It is suggested by [17] that local rate information affects phonetic categorization more strongly than habitual rate information since normalization for local rate information occurs too early during speech perception for it to be influenced by habitual rate information. If listeners primarily made use of the invariable local rate information, it is then unsurprising that no rate effect was found. However, previous studies found that the speech rate of preceding sentential context of a target stop reliably modulates VOT perception independent of the duration of the post-stop vowel [7, 25, 26]. Therefore, the lack of habitual speech rate effect in our study cannot be attributed to the absolute dominance of the post-stop vowel length cue over preceding contextual speech rate cues in VOT perception.

Another possibility is that listeners did not make use of the habitual speech rate of the speakers in VOT perception due to high variability across speakers in VOT realization [1]. For example, as speakers age, speech rate slows down and vowels are produced longer but VOT values become shorter [4]. Given this interspeaker variability, a faster speech rate of one speaker is not a reliable indicator that they will

produce a short VOT compared to a slower speaker. Moreover, the English VOT contrast may be robust enough against speaker-specific speech rate variation and a rate-independent VOT boundary is effective enough for English voicing categorization in conversational speech, obviating the need for rate normalization [16].

Further research is required to examine if habitual speech rate effects on subsequent sound categorization are contingent upon the strength of interspeaker consistency in correlation with general speech rate and the particular durational contrasts.

5. ACKNOWLEDGEMENTS

This research was supported by the University of Toronto Scarborough Research Competitiveness Fund. Thanks to Na-Young Ryu and Hyoung-Seok Kwon for their assistance in the implementation of on-line experiments.

6. REFERENCES

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

[2] 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.

[3] Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Grothendieck, G. 2017. lme4 package, version 1.1-13 [Comupter software]. Available from https://cran.r-project.org/web/packages/lme4/

[4] Benjamin, B. J. 1982. Phonological performance in gerontological speech. *Journal of Psycholinguistic Research* 11(2), 159-167.

[5] Boersma, P., Weenink, D. 2017. *Praat: Doing phonetics by computer* (Version 6.0.34) [Computer program]. Retrieved from http://www.praat.org/

[6] Gay, T. 1978. Effect of speaking rate on vowel formant movements. *J. Acoust. Soc. Am.* 63, 223–230.

[7] Kang, Y., Kung, K., Li, J., Ting, C., Yeung, J. 2018. Compensating for speech rate in English stop perception. *Toronto Working Papers in Linguistics* 40.

[8] Kessinger, R. H., Blumstein, S. E. 1997. Effects of speaking rate on voice-onset time in Thai, French, and English. *J. Phon.* 25, 143–168.

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

[10] Maslowski, M., Meyer, A. S., Bosker, H. R. 2018. How the Tracking of Habitual Rate Influences Speech Perception. *J. Exp. Psychol. Learn. Mem. Cogn*. Advance online publication. http://dx.doi.org/10.1037/xlm0000579

[11] 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.

[12] 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.

[13] Miller, J. L., Liberman, A. M. 1979. Some effects of later-occurring information on the perception of stop consonant and semivowel. *Perception & Psychophysics* 25, 457–465.

[14] Miller, J. L., Volaitis, L. E. 1989. Effect of speaking rate on the perceptual structure of a phonetic category. *Percept and Psychophys* 46, 505–512.

[15] Miller, J. L., Green, K. P., Reeves, A. 1986. Speaking rate and segments: A look at the relation between speech production and speech perception for the voicing contrast. *Phonetica* 43(1–3), 106–115.

[16] Nakai, S., Scobbie, J. M. 2016. The VOT Category Boundary in Word-Initial Stops: Counter-Evidence Against Rate Normalization in English Spontaneous Speech. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 7(1), 13.

[17] 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.

[18] Port, R. F. 1981. Linguistic timing factors in combination. *J. Acoust. Soc. Am.* 69, 262-274.

[19] R development Core Team. 2016. R: A Language and Environment for Statistical Computing. Version 3.3.2. Vienna, Austria: R Foundation for Statistical Computing.

[20] Reinisch, E., Sjerps, M. J. 2013. Compensation for speaking rate and spectral context take place at a similar point in time. *J. Phon.* 41, 101–116.

[21] Reinisch, E. 2015. Speaker-specific processing and local context information: The case of speaking rate. *Applied Psycholinguistics* 37, 1397-1415.

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

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

[24] Theodore, R. M., Miller, J. L., DeSteno, D. 2009. Individual talker differences in voice-onset-time; Contextual influences. *J. Acoust. Soc. Am.* 125, 3974–3982.

[25] Toscano J. C., McMurray B. Cue integration and context effects in speech: Evidence against speaking rate normalization. Attn Percep Psychphys. 2012; 74:1284–1301.

[26] 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.

[27] Volaitis, L. E., Miller, J. L. 1992. Phonetic prototypes: Influence of place of articulation and speaking rate on the internal structure of voicing categories. *J. Acoust. Soc. Am.* 92(2), 723–735.