Speaking rate normalization across different talkers in the perception of Japanese stop and vowel length contrasts

Misaki Kato[1], *Shigeto Kawahara*[2], *Kaori Idemaru*[3]

[1,3] University of Oregon, Eugene, Oregon, U.S.A
[2] Keio University, Tokyo, Japan
misaki@uoregon.edu, kawahara@icl.keio.ac.jp, idemaru@uoregon.edu

Abstract

Perception of duration is critically influenced by the speaking rate of the surrounding context. However, to what extent this speaking rate normalization depends on a specific talker's voice is still understudied. The present study investigated whether listeners' perception of temporally contrastive phonemes is influenced by the speaking rate of the surrounding context, and more importantly, whether the effect of the contextual speaking rate persists across different talkers for different types of contrasts: Japanese singleton-geminate stop contrast (/k/-/kk/) and short-long vowel contrast (/e/-/ee/). The vowel contrast carries more reliable talker information than the stop contrast; hence, listeners' rate-based adjustments may be more talker-specific for vowels than for stops. The current results showed that context speaking rate impacted the perception of the target contrast across different talkers, and this influence was evident for both types of the contrasts tested. These results suggest that listeners generalized their rate-based adjustments to different talkers' speech regardless of whether the target segment carried reliable talker information (i.e., vowel contrast) or not (i.e., stop contrast). The current results bear on the issue of how speaking rate information is processed with respect to talker information.

Index Terms: speech perception, speaking rate, length contrast, categorical perception, Japanese

Introduction

Even within a single language, different people speak differently. One aspect of speech in which talkers vary significantly is how fast they speak; some people talk faster than others [1,2] and the same person may talk faster or slower in different occasions [3]. It has been demonstrated that listeners take this speaking rate variation into account when processing speech [4-7]. One piece of evidence for this rate-dependent speech perception is observed as the phonetic boundary shift in listeners' perception of temporally contrastive phonemes. For example, the perceptual boundary between English /b/ vs. /p/—characterized by different VOT durations—changes depending on the speaking rate of the surrounding speech [8-10]. More specifically, perception of the English /p/-/b/ continuum (as in *rapid* vs. *rabid*) is biased toward /p/ when the target word including these sounds follows a faster than a slower precursor phrase [11]. Similar rate effects have been found in perception of other contrasts involving temporal cues [12,13], manner of articulation [14], lexical stress [5], word segmentation [15], as well as in the perception of function words [16,17].

One question that arises is how strongly listeners' rate-based normalization is associated with a specific talker's voice. That is, is listeners' auditory normalization of phonetic temporal cues based on general auditory input (e.g., speech produced by multiple talkers) or is it based on the speech produced by a specific talker? Some previous studies have suggested that listeners track talker information that is carried in the acoustic signal, store this information along with sound or word representations, and use this information when processing new speech [18-21]. Given such reports, it is plausible that listeners' perceptual learning—and resulting perceptual processes—are attuned to specific talkers. This prediction is supported by the studies demonstrating that listeners' perceptual learning of phoneme categories is talker-specific [22,23]. Specifically, these studies have shown that when listeners are exposed to a particular talker's speech, they adjust their phonemic categories for that specific talker, but do not generalize the adjustments to a different talker. However, other results support a different view, demonstrating that listeners generalize their perceptual learning of phonemic categories based on one talker to a different talker [24]. Further, it has been shown that the speaking rate of one talker affects the perception of another talker [10,25]. That is, the speaking rate of the context surrounding the critical segment affects the perception of the critical segment even if the context is produced by a different talker than that of the critical segment. In short, the previous results are mixed regarding whether listeners' perceptual adjustments of phonetic boundaries are talker-specific or not.

It is possible that these apparently contradicting findings are due to specific acoustic characteristics of the segments that were tested in these studies. Kraljic and Samuel [26] have demonstrated that listeners' perceptual adjustments of a phonemic category boundary differed depending on the sound contrast that is being varied (i.e., /d/-/t/ vs. /s/-/ʃ/). They suggested that when acoustic cues that differentiate the target contrast simultaneously provide information about the talker's identity (i.e., fricatives), listeners' perceptual adjustment depends on the talker, while the adjustment is independent of talkers when the acoustic cues are less informative for identifying the talker (i.e., stops). Given this finding, it is possible that listeners' rate-based adjustments for temporally contrastive phonemes also differ depending on whether the target segment carries reliable talker information in addition to the temporal cues that distinguish the target contrast itself. That is, listeners' rate-based adjustments may not generalize to different talkers' speech (i.e., talker-dependent adjustment) when the acoustic signal of the temporally contrastive phonemes also carries reliable talker information, whereas the adjustments may generalize across talkers (i.e., talker-independent adjustment) when the acoustic signal of the contrast does not carry talker information.

The current study investigates the effect of speaking rate variation of the surrounding context on the perception of temporally contrastive phonemes. Specifically, we examine whether the speaking rate of a precursor phrase produced by one talker impacts the perception of the target contrast produced by a different talker, and if this pattern differs depending on whether or not the target segment carries talker information. We examine this question using two different types of contrasts in Japanese. Japanese has a singleton-geminate contrast (e.g., /k/-/kk/) as well as a short-long vowel contrast (e.g., /e/-/ee/), both of which are primarily based on durational differences [27]. These contrasts differ in terms of whether the acoustic characteristics of the contrast also carry reliable talker information, including gender differences. Particularly, while the difference between male vs. female voice is carried in the difference in spectral characteristics of the vowels, it is manifested much less clearly in the closure/silence intervals of stops. Thus, it is possible that the speaking rate of the precursor phrase produced by one talker affects the perception of a stop contrast produced by a different talker, while the same would not hold for the perception of a vowel contrast. However, it is also possible that listeners' rate-based adjustments generalize to different talkers' speech regardless of the type of the target segment (i.e., stops and vowels), because both contrasts are primarily duration-based. That is, unlike the English /s/-/ʃ/ contrast that differs in the spectral dimension, which also varies with the gender of the talker [26], the Japanese singleton-geminate stop contrast (/k/-/kk/) and short-long vowel contrast (/e/-/ee/) both differ in the temporal dimension, which is much less directly related to the gender of the talker compared to spectral differences. Thus, the speaking rate variation of the precursor phrase may affect the perception of the target contrast even if the precursor phrase and the target word are produced by different talkers, and this pattern may persist for both types of segments (i.e., stop consonant and vowel contrasts).

Methods

1.1. Participants

Participants were 15 native Japanese listeners (11 females, 4 males; age mean = 21.2 years, range = 20-26 years), who were residing in the US at the time of testing. They were all familiar with English as their second language. None of them reported a history of speech or hearing impairment.

1.2. Materials

The precursor phrase was */kikoeta-kotoba-wa/* ("*the word I heard was* ___"). The target segments, stop consonant and vowel, were embedded word-medially in non-words: /he**k**o-he**kk**o/ (consonant) and /h**e**su-h**ee**su/ (vowel). Two native Japanese talkers (1 female, 1 male) recorded multiple tokens of the precursor phrase and both singleton and geminate versions of the target words. The talkers were residing in the US at the time of recording, and all spoke the standard Japanese. In a sound booth, the materials were displayed on the computer screen one at a time; the presentation was self-paced. The speech was recorded using a microphone that was directly connected to a desktop computer, using a mono channel at a sampling rate of 44,100 Hz (16 bit) using the Praat speech analysis software package [28]. The target words were produced with a high-low or high-low-low pitch pattern (i.e. with initial pitch accent, the default accent pattern for nonce words). The clearest tokens of the precursor phrase and target words were chosen from each talker.

The durations of the precursor phrase and segments in the target words were adjusted using the Pitch Synchronous Overlap and Add (PSLOA) algorithm in Praat. Specifically, the two talkers used for this study were selected from the pool of six talkers who all provided the speech materials; the selection was made based on the results of a pilot study examining the clarity of their productions. The precursor and target word durations of the two selected talkers were adjusted to be the mean durations across the six talkers. This mean duration of the precursor phrase was further manipulated through linear expansion (factor of 1.6) and linear compression (factor of 1/1.6 = .625) with PSOLA, resulting in three rates: fast, normal (no further rate manipulation), and slow. These precursor phrases were RMS normalized to 75 dB. To create target word continua, the duration of the target segments (i.e., /k/ in /heko-hekko/ and /e/ in /hesu-hessu/) were manipulated in five 20 ms steps (i.e., 60, 80, 100, 120, 140 ms) so that the range encompasses typical short and long segments [29]. The target words were then RMS normalized to 70 dB.

Finally, the precursor phrase and target words were concatenated so that all precursors (3 rates x 2 talkers) were combined with all target words (2 segments x 5 durations x 2 talkers), resulting in 120 unique stimuli. Congruent stimuli were those in which the voice of the precursor and the target matched, and incongruent stimuli were those in which the precursor and target voices did not match.

1.3. Procedure

Participants were seated in front of a computer wearing headphones in a sound-attenuated room. A forced-choice perception experiment was delivered via Psychopy [30]. In each trial, participants heard a sentence through the headphones, simultaneously saw two response choices (e.g., /heko/ and /hekko/) in Japanese orthography on the screen, and were asked to choose the word they heard by pressing the key 'f' (short: /heko/ or /hesu/) or 'j' (long: /hekko/ or /heesu/). They were instructed to respond as quickly and accurately as possible. Consonant and vowel trials were blocked, and the order of the two blocks were counter-balanced across participants. Within each block, there were two practice trials preceding the test trials, and the test stimuli (i.e., 60 consonant stimuli and 60 vowel stimuli) were presented to each participant in 5 randomized orders. The entire session lasted approximately 45 minutes.

1.4. Analysis

Responses were analyzed using mixed-effects logistic regression models using R package lme4 [31] where the short (/heko/ or /hesu/) or long response (/hekko/ or /heesu/) was the dependent variable. As shown in the model syntax below, the fixed factors included target segment duration (centered, continuous), condition (categorical: congruent or incongruent), precursor rate (categorical: fast, normal, or slow), segment (categorical: consonant or vowel), and interactions of these factors. Each categorical fixed factor was treatment-coded; the reference level (i.e., the level coded as 0) for segment was consonant, for precursor rate was normal, for condition was congruent. We were interested in examining whether the effect of the precursor rate persists when the precursor voice and the target voice are different (incongruent condition) and whether this effect was present for both vowel and consonant contrasts (segment type). Thus, our main interest was the three-way interaction among precursor rate, condition, and segment type. We excluded the three-way interaction among duration, condition, and precursor rate as well as the four-way interaction because these interactions were not relevant to our research questions, and also to avoid convergence problems. The maximal random effects structure that would converge was implemented, which included random intercepts for listener, as well as by-listener random slopes for segment duration, condition, precursor rate, segment, and the interaction between condition and segment, and between precursor rate and segment. We uncorrelated random factors to aide convergence problems.

```
Response ~ Duration* Condition* Rate* Segment
                  - Duration: Condition: Rate
                  - Duration: Condition: Rate: Segment +
(1+ Duration+ Condition*Segment+ Rate*Segment || Listener)
```

Results

Figure 1 illustrates the proportion of long responses (i.e., /hekko/ or /heesu/) across the 5 steps of the target segment duration continua and 3 levels of precursor rate (fast, normal, or slow) by condition (congruent or incongruent). The summary of the mixed-effects logistic regression model is shown in Table 1 at the end of the paper. The results of the model showed a significant effect of duration ($\beta = .11$, $z = 12.11$, $p < .001$), and the effect of duration interacted with segment ($\beta = -.02$, $z = -3.19$, $p < .01$), indicating that the 'long' responses increased along the 5-step duration continua but this increase was smaller for vowel (as shown in Figure 1).

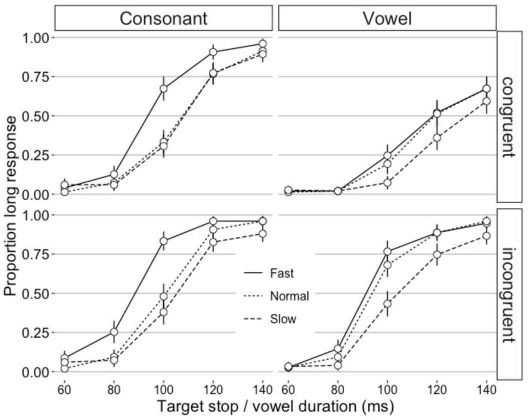

Figure 1: *Proportion of long responses for two segments (consonant /hekko/ or vowel /heesu/) for 3 levels of precursor rate (fast, normal, slow) across 5 steps of target stop closure or vowel duration (ms) by condition (congruent or incongruent). Error bars indicate the 95% confidence interval of the mean.*

Figure 2 is a different illustration of the same results. It collapses the duration continua from Figure 1 (i.e., x-axis in Figure 1), illustrating the effects of segment, condition, and precursor rate. The effect of precursor rate was significant for the normal vs. fast comparison ($\beta = 1.32$, $z = 7.18$, $p < .001$), but not for the normal vs. slow comparison ($\beta = .1$, $z = .57$, $p = .57$). These effects of precursor rate interacted with segment (normal vs. fast x segment: $\beta = -1.11$, $z = -4.31$, $p < .001$; normal vs. slow x segment: $\beta = -.57$, $z = -2.2$, $p < .05$). These results indicate that the effects of precursor rates on listeners' perception of the target segment differed across different segments (consonant vs. vowel). In order to further examine these interactions between precursor rate and segment, a post-hoc test assessed the effects of precursor rates separately for consonant and vowel. The results showed that for consonant, the fast vs. slow and normal vs. fast comparisons were significant ($p < .0001$ for both comparisons), but not normal vs. slow ($p = .37$). For vowel, the fast vs. slow and normal vs. slow comparisons were significant ($p < .0001$ for both), but not normal vs. fast ($p = .19$). These results indicate that although the precursor rate effects were present for both consonant and vowel, the source of difference varied slightly. The fast vs. slow difference influenced perception for both consonant and vowel. Consonant perception was further affected by the *normal vs. fast* difference, though vowel perception was affected by the *normal vs. slow* difference.

These patterns of rate*segment interactions did not differ across congruent vs. incongruent conditions as indicated by the non-significant three-way interactions among segment, condition, and normal vs. fast precursor rate ($\beta = -.035$, $z = -.1$, $p = .92$), and among segment, condition, and normal vs. slow precursor rate ($\beta = -.33$, $z = -.1$, $p = .92$). In terms of the effects of rate and condition (congruent vs, incongruent), there was a significant interaction between the normal vs. slow comparison and condition ($\beta = -.55$, $z = -2.36$, $p < .05$), but not between the normal vs. fast comparison and condition ($\beta = .09$, $z = .37$, $p = .71$). Post-hoc tests revealed that the normal vs. fast comparison was significant in both congruent and incongruent conditions ($p < .0001$ for both). However, the normal vs. slow comparison was significant in the incongruent ($p < .0001$), but not in the congruent condition ($p = .38$). This was likely affected by the pattern that the normal vs. slow difference in precursor rate did not affect consonant perception (as discussed above), and this was especially the case in the congruent condition.

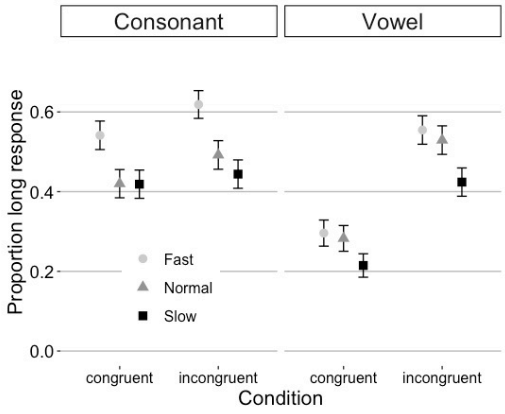

Figure 2: *Proportion of long responses for two segments (consonant /hekko/ or vowel /heesu/) for 3 levels of precursor rate (fast, normal, slow) by condition (congruent or incongruent). Error bars indicate the 95% confidence interval of the mean.*

Further, as shown in the figures, the proportion of long responses for vowel was lower in the congruent than in the incongruent condition. This was reflected in several significant terms in the model, including the effect of condition ($p < .01$), segment ($p < .01$), the interaction between condition and segment ($p < .001$), as well as the three-way interaction among duration, condition, and segment ($p < .01$).

Overall, these results demonstrate that the effect of precursor rates (fast, normal, slow) was present in both congruent and incongruent conditions (i.e., when the voice of the precursor and target match and mismatch) and in both segments (i.e., consonant and vowel). The way the precursor rates affected listeners' perception of the temporal contrast was slightly different across the target segments.

Discussion

The present study investigated whether listeners' perception of temporally contrastive phonemes is influenced by the speaking rate of the precursor phrase when the talker of the precursor and the target word match (congruent) and mismatch (incongruent), and whether this pattern differs for different target contrasts: Japanese singleton-geminate stop contrast (i.e., /k/-/kk/) and short-long vowel contrast (i.e., /e/-/ee/). The results demonstrated that the effect of precursor rates was present in both congruent and incongruent conditions, and this pattern persisted for both contrasts. That is, the faster the precursor rate was, the more often the target phoneme was perceived as the 'long' phoneme (i.e., geminate stop /kk/ and long vowel /ee/) even when the talker of the precursor phrase differed from that of the target word. This general effect of the precursor rate manifested itself somewhat differently for the consonant and vowel contrasts; the effect of the normal vs. fast difference impacted the perception of the consonant contrast, but the normal vs. slow difference impacted the perception of the vowel contrast. However, since the listeners were not explicitly made aware of the *normal* precursor rate as the reference during the experiment, it is possible that their rate-based adjustments were made more broadly than based on the specific comparisons of how *fast* and *slow* speaking rates deviated from the *normal* rate. In fact, the *fast* vs. *slow* difference affected the perception of both consonant and vowel contrasts, indicating that speaking rate variation of the precursor phrase generally impacted the perception of the two target contrasts. Together, these results suggest that listeners generalized their rate-based adjustments of the target contrast to different talkers' speech both when the target segment carried reliable talker information (i.e., vowel contrast) and when it did not (i.e., stop contrast).

The present results appear to suggest that listeners' rate-based adjustments are independent of talkers. Listeners adjusted their perception of temporally cued segments (short vs. long consonants and vowels) using the speaking rate of the surrounding context even when the context was spoken in a different voice than that of the critical segment. This is in line with the previous results suggesting that rate normalization is an obligatory process, where listeners use any available information to make rate-based adjustments [9,10,25]. Studies have shown that listeners' rate-based adjustments are robust even for the irrelevant talker's voice presented simultaneously with the relevant talker's voice [25] and under conditions with varying attentional demands [32]. The present results contribute to these lines of research demonstrating that rate normalization across talkers persists even when the target segment reliably signals the talker difference (i.e., vowels). The current result may be taken as further evidence for the claim that rate-based speech perception is governed by general auditory normalization processes that occur early in perception [10,14,32,33]. That is, extraction of rate information may occur earlier than segregation of voices, and the rate information affects subsequent auditory processing.

While these results suggest that rate-based adjustments operate regardless of talker information, it is possible that listeners in the current study may have been inclined to disregard talker information, because the precursor phrases of the two voices had the same durations (i.e., the durations of the fast, normal, and slow rates were the same across the two voices). Thus, listeners may have been more focused on adjusting their perception for different rates rather than for different talkers. However, listeners may be more sensitive to talker information when processing rate information that carries within-talker variation specific to a particular talker. For example, studies have shown that listeners keep track of different talkers' habitual (global) speaking rates, as opposed to the rate of the local context (i.e., phrases immediately preceding the target contrast as in the present study) [34,35]. Given these results, it is possible that listeners' rate-based adjustments for different segments (consonants vs. vowels) may show different patterns if listeners are exposed to the habitual rate of different talkers. That is, talker-independent rate adjustment may be more robust in stop length contrasts than vowel length contrasts when listeners are more familiar with different talkers' habitual rates as compared to just local context rates. Additionally, vowel and stop length contrasts may differ not only in the amount of talker information carried in the segment but also with respect to other factors (e.g., perceptual salience of duration information). Further investigation should thus address the nature of generality and specificity of rate-based perception in relation with the type of phenetic environment that carries the information.

Table 1: *Summary of the mixed-effects logistic regression model.*

	Est.	**S.E.**	***z* val.**	***p***
(Intercept)	-.89	.26	-3.42	.000
Duration	.11	.01	12.11	.000***
Condition	.79	.25	3.12	.002**
Rate [Fast]	1.32	.18	7.18	.000***
Rate [Slow]	.10	.17	.57	.57
Segment	-1.11	.36	-3.06	.002**
Duration: Condition	.002	.005	.39	.69
Duration: Rate [Fast]	.0001	.007	.02	.99
Duration: Rate [Slow]	-.02	.006	-3.19	.001**
Condition: Rate [Fast]	.09	.25	.37	.71
Condition: Rate [Slow]	-.55	.23	-2.36	.018*
Duration: Segment	-.02	.007	-3.19	.001**
Condition: Segment	1.62	.3	5.5	.000***
Rate [Fast]: Segment	-1.11	.26	-4.31	.000***
Rate [Slow]: Segment	-.57	.26	-2.2	.028*
Dur: Cond: Seg	.02	.007	3.09	.002**
Dur: Rate [Fast]: Seg	-.005	.009	-.59	.56
Dur: Rate [Slow]: Seg	.005	.008	.56	.58
Cond: Rate [Fast]: Seg	-.035	.36	-.1	.92
Cond: Rate [Slow]: Seg	-.033	.34	-.1	.92

5. References

[1] T. H. Crystal and A. S. House, “Segmental durations in connected speech signals: Current results,” *The Journal of the Acoustical Society of America*, vol. 83, pp. 1553–1573, 1988.

[2] H. Quené, “Multilevel modeling of between-speaker and within-speaker variation in spontaneous speech tempo,” *The Journal of the Acoustical Society of America*, vol. 123, no. 2, pp. 1104-1113, 2008.

[3] J. L. Miller, F. Grosjean, and C. Lomanto, “Articulation rate and its variability in spontaneous speech: A reanalysis and some implications,” *Phonetica,* vol. 41, pp. 215–225, 1984.

[4] H. R. Bosker, “Accounting for rate-dependent category boundary shifts in speech perception,” *Attention, Perception, & Psychophysics*, vol. 79, no. 1, pp. 333-343, 2017.

[5] E. Reinisch, A. Jesse, and J. M. McQueen, “Speaking rate from proximal and distal contexts is used during word segmentation,” *Journal of Experimental Psychology: Human Perception and Performance*, vol. 37, no. 3, pp. 978 – 996, 2011a.

[6] Q. Summerfield, “Articulatory rate and perceptual constancy in phonetic perception,” *Journal of Experimental Psychology: Human Perception and Performance*, vol. 7, no. 5, pp. 1074–1095, 1981.

[7] S. C. Wayland, J. L. Miller, and L. E. Volaitis, “The influence of sentential speaking rate on the internal structure of phonetic categories,” *The Journal of the Acoustical Society of America*, vol. 95, no.5, pp. 2694-2701, 1994.

[8] G. R. Kidd, “Articulatory-rate context effects in phoneme identification. *Journal of Experimental Psychology: Human Perception and Performance,”* vol. 15, no. 4, pp. 736-748, 1989.

[9] J. L. Miller and E. R. Dexter, “Effects of speaking rate and lexical status on phonetic perception,” *Journal of Experimental Psychology: Human Perception and Performance,* vol. 14, no. 3, pp. 369-378, 1988.

[10] J. R. Sawusch and R. S. Newman, “Perceptual normalization for speaking rate II: Effects of signal discontinuities,” *Perception & psychophysics*, vol. 62, no. 2, pp. 285-300, 2000.

[11] P.C. Gordon, “Induction of rate-dependent processing by coarse-grained aspects of speech,” *Perception & Psychophysics*, vol. 43, no.2, pp. 137-146, 1988.

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

[13] E. Reinisch and M. J. Sjerps, “The uptake of spectral and temporal cues in vowel perception is rapidly influenced by context,” *Journal of Phonetics*, vol. 41, pp. 101–116, 2013.

[14] T. Wade and L. L. Holt, “Perceptual effects of preceding nonspeech rate on temporal properties of speech categories,” *Perception & Psychophysics,* vol. 67, pp. 939–950, 2005.

[15] E. Reinisch, A. Jesse, and J. M. McQueen, “Speaking rate affects the perception of duration as a suprasegmental lexical-stress cue,” *Language and Speech,* vol. 54, pp. 147–166, 2011b.

[16] L. C. Dilley and M. A. Pitt, “Altering context speech rate can cause words to appear or disappear,” *Psychological Science*, vol. 21, pp. 1664–1670, 2010.

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

[18] S. C. Creel, R. N. Aslin, and M. K. Tanenhaus, “Heeding the voice of experience: The role of talker variation in lexical access,” *Cognition*, vol. 106, no. 2, pp. 633-664, 2008.

[19] S. C. Creel and M. R. Bregman, “How talker identity relates to language processing,” *Language and Linguistics Compass*, vol. 5, no. 5, pp. 190-204, 2011.

[20] S. D. Goldinger, “Words and voices: episodic traces in spoken word identification and recognition memory,” *Journal of experimental psychology: Learning, memory, and cognition*, vol. 22. no. 5, pp. 1166-1183, 1996.

[21] L. C. Nygaard and D. B. Pisoni, “Talker-specific learning in speech perception,” *Perception & psychophysics*, vol. 60, no. 3, pp. 355-376. 1998.

[22] F. Eisner and J. M. McQueen, “The specificity of perceptual learning in speech processing,” *Perception & psychophysics*, vol. 67, no. 2, pp. 224-238, 2005.

[23] T. Kraljic and A. G. Samuel, “Perceptual learning for speech: Is there a return to normal?” *Cognitive psychology*, vol. 51, no. 2, pp. 141-178, 2005.

[24] T. Kraljic and A. G. Samuel, “Generalization in perceptual learning for speech,” *Psychonomic bulletin & review*, vol. 13, no. 2, pp. 262-268, 2006.

[25] R. S. Newman, J. R. Sawusch, J. R. “Perceptual normalization for speaking rate III: Effects of the rate of one voice on perception of another,” *Journal of phonetics*, vol. 37, no. 1, pp. 46-65, 2009.

[26] T. Kraljic, and A. G. Samuel, “Perceptual adjustments to multiple speakers,” *Journal of Memory and Language*, vol. 56, no. 1, pp. 1-15, 2007.

[27] T. J. Vance, *The sounds of Japanese with audio CD*. Cambridge: Cambridge University Press, 2008.

[28] P. Boersma and D. Weenink, “Praat: Doing phonetics by computer” [Computer program], 2015.

[29] S. Kawahara, “The phonetics of *sokuon*, obstruent geminates,” in *Handbook of Japanese Language and Linguistics*. Berlin: De Gruyter Mouton, 2015, pp. 43-73.

[30] J. W. Peirce, “PsychoPy—Psychophysics software in Python,” *Journal of Neuroscience Methods,* vol. 162, no. 1-2, pp. 8-13, 2007.

[31] D. Bates, M. Mächler, B. Bolker, and S. Walker, “Fitting linear mixed-effects models using lme4,” *Journal of Statistical Software*, vol. 67, pp. 1–48, 2015.

[32] H. R. Bosker, E. Reinisch, and M. J. Sjerps, “Cognitive load makes speech sound fast, but does not modulate acoustic context effects,” *Journal of Memory and Language*, vol. 94, pp. 166-176, 2017.

[33] J. Kingston, S. Kawahara, D. Chambless, D. Mash, and E. Brenner-Alsop, “Contextual effects on the perception of duration,” *Journal of Phonetics*, vol. 37, no. 3, pp. 297-320, 2009.

[34] E. Reinisch, “Speaker-specific processing and local context information: The case of speaking rate,” *Applied Psycholinguistics*, vol. 37, no. 6, pp. 1397-1415, 2016.

[35] M. Maslowski, A. S. Meyer, and H. R. Bosker, “How the tracking of habitual rate influences speech perception,” *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 45, no. 1, pp. 128-138, 2019.