# Speech rate and the perception of consonant and vowel length in Japanese: neural entrainment and episodic memory

Timothy Gadanidis,[1] Yoonjung Kang,[2,1,a]

[1] *Department of Linguistics, University of Toronto, Toronto, ON, Canada*

[2] *Department of Language Studies, University of Toronto Scarborough, Toronto, ON, Canada*

*timothy.gadanidis@mail.utoronto.ca*, *yoonjung.kang@utoronto.ca*

[a] Author to whom correspondence should be addressed.

Neural entrainment (persistence of neural oscillation that captures speech rhythm) and episodic memory-based perception (exemplar theory and belief-updating model) have both been proposed to account for speech rate-dependent perception. This paper probes these models by testing rate-dependent perception of Japanese vowel and stop length contrasts. Contextual speech rate was varied by manipulating all segments, vowels only, or consonants only in the carrier sentence. Robust rate effects were found across the board, supporting the neural entrainment account, and partial support was found for the episodic account: the rate effect was modulated by manipulation method for stops, but not for vowels.

## 1. Introduction

Speech is highly variable. Some sources of this variation are random, but many are structured and systematic, such as the effects of speakers' physiological or social differences and linguistic factors such as prosody, adjacent segments, and speech style. At issue in the present study is variation by speech rate: as the rate of speech increases or decreases, segments are shortened, or lengthened. For durationally-signaled contrasts, shortening and lengthening of segments requires rate-dependent speech perception. Listeners have been shown to take into account the rate of speech when identifying the length of a given stimulus — identical stimuli are more likely to be heard as 'long' when embedded in fast speech and 'short' in slow speech (Miller, Green & Reeves, 1986), and the rate effect is found even when the rate cue is distal, i.e., in the same utterance but one or more syllables away (Reinisch et al., 2011, Heffner et al., 2017).

Neural entrainment has been proposed as the cause of rate-dependent perception. Kösem et al. (2018) demonstrate that neural oscillation tracks the speech rhythm, and the oscillation persists a few cycles past stimulation (i.e., the speech rate cue), shaping the perception of temporal contrasts in subsequent speech. However, neural entrainment is unlikely to be the only mechanism at play in rate normalization: it cannot explain context-dependent perception involving non-temporal cues (e.g., vowel quality, fricative place, and pitch) and cannot stand as a general account of context-dependent perception. Also, it has been shown that listeners can track speaker-specific speech rates and use them in subsequent perception distinctly

41 (Maslowski et al., 2019, Reinisch, 2016), which is not straightforward to explain
42 under the neural entrainment account.
43 On the other hand, models have been proposed to account for listeners'
44 ability to perceive the target structure accurately despite variability along diverse
45 acoustic dimensions and linguistic and non-linguistic contexts. The exemplar-based
46 model of speech perception (Pierrehumbert, 2001) holds that speakers store rich
47 acoustic detail about past production of a structure so that speaker- and situation-
48 specific information can be used in later perception. The belief-updating model of
49 speech adaptation (Kleinschmidt and Jaeger, 2015) holds that when listeners
50 recognize consistencies in the speech signal for a given situation, they will track
51 situation- or speaker-specific distributions of acoustic cues. Listeners apply their
52 knowledge of these context-specific cue distributions in perception when the context
53 is recognized again. In these models, rate normalization arises when listeners infer
54 whether a segment they heard is 'long' or 'short', not based on the absolute duration,
55 but based on their knowledge of speech rate-specific distributions of the long and
56 short segments.
57 In this study, we tested these accounts of rate-dependent perception by
58 investigating whether the category-specific durational distributions in the ambient
59 speech affect subsequent length perception in Japanese. Japanese is an ideal
60 environment for testing multiple durational contrasts because it has a phonemic
61 length distinction for both (voiceless) stops and vowels (e.g., kito = 'returning', kitto
62 = 'surely', kiito = 'raw silk thread'). Both of these contrasts are potentially affected
63 by speech rate variation: the durations of long vowels at fast speech rates overlap
64 with the durations of short vowels at slow speech rates (Hirata, 2004), and the same

65 is true for stops (Hirata and Whiton, 2005). This allows us to probe how different
66 segment types are affected by rate normalization when we vary the way fast and slow
67 speech rates in the ambient speech are achieved, namely by either altering the
68 duration of all segments equally or by manipulating only one category of segments. If
69 neural entrainment is the sole or primary mechanism of rate normalization, we may
70 see little or no difference based on which segments are manipulated. On the other
71 hand, if rate normalization is due to context-specific inference of target structure, we
72 expect different rate normalization patterns depending on the category-specific
73 durational cues.

74 We report the results of five perception experiments on length perception of
75 Japanese vowels and stops under varying speech rate conditions. We first confirm
76 that rate normalization does occur for both categories (Experiment 1). We then find
77 that how the rate is manipulated affects the perception of stops but not vowels: we
78 find a weaker rate effect for stops when rate is varied by manipulating vowels only
79 rather than all segments equally, while no difference was found for vowels
80 (Experiment 2). This asymmetry holds even when the pitch cue for length is
81 removed, which is a strong cue for vowel length but not for stop length (Experiment
82 3) or when the manipulation conditions are presented in separate blocks
83 (Experiment 4). The consonant-vowel asymmetry also holds in a consonant-only
84 manipulation condition: the rate effect for stops is stronger when the rate is changed
85 by manipulating consonants only rather than all segments equally, but no difference
86 is found for vowels (Experiment 5).

87 **2. Methods**

We built five online experiments using *jsPsych* (de Leeuw, 2015) and recruited participants from prefectures throughout Japan via CrowdWorks, a Japanese crowdsourcing site. The gender breakdown (F:M) of the participants for the five experiments was 13:17, 16:24, 20:20, 20:19, and 8:31+11, respectively. The range and the median of the participants' age were comparable across the experiments: Exp. 1: 27-62 (37.5), Exp. 2: 27-64 (42), Exp. 3: 20–62 (40), Exp. 4: 21–65 (42), and Exp. 5: 24–61 (40). The source code for the experiments as well as the data, manipulation and analysis scripts, and stimuli are provided in the Data Availability Statement. We created duration continua based on four vowel-length minimal pairs (/kádo/ 'corner' ~ /ká:do/ 'card'; /bíru/ 'building' ~ /bí:ru/ 'beer'; /médo/ 'goal' ~ /mé:do/ 'brightness'; /tóʃo/ 'books' ~ /tó:ʃo/ 'beginning') and four consonant-length minimal pairs (/méta/ 'meta' ~ /mét:a/ 'thoughtless'; /táte/ 'height' ~ /tát:e/ 'stand up'; /káko/ 'the past' ~ /kák:o/ 'parentheses'; /íka/ 'below' ~ /ík:a/ 'family'), some of which were drawn from Hirata (2004).

A male native speaker of Kanto Japanese produced both members of each target pair embedded in a carrier sentence, which was designed not to include any long segments: *Takeuchi-san wa totemo odayaka ni [word] to hatsuonshita* "Mr. Takeuchi very calmly pronounced [word]." One production of the carrier sentence was chosen and all target words were spliced into it at the same location to create the final versions of each baseline stimulus. Then, the duration continua were created from tokens of both a long and short member of each pair using Praat's PSOLA algorithm (Boersma and Weenink, 2021). The duration steps for each pair of continua were chosen to make estimated responses comparable across all word pairs based on the results of a 10-listener pilot test, also conducted via CrowdWorks: the midpoint of

each continuum was the point at which the odds of a participant in the pilot test hearing it as "short" or "long" were predicted to be even based on a regression model, with the ends of each continuum being the points at which the response was predicted to approach near-100% short and near-100% long. For each continuum, we manipulated the duration of the carrier sentence to be either 20% shorter ("fast") or 20% longer ("slow") than the original production. Participants listened to the stimuli embedded in a carrier sentence and selected the normal orthographic form of the word they heard by key press. Except for Experiment 4, items were presented in a pseudorandom order.

Statistical analysis was conducted using R (R Core Team, 2021). Mixed-effects logistic regression models were fitted using the glmer() function from the *lme4* package (Bates et al., 2015b). Each model was built to predict the frequency of a 'long' (1) response as opposed to a 'short' (0) response. Figures were created using the *ggplot2* package (Wickham, 2016), using the *pilot* theme (Hawkins 2024). The predictors used in the experiments are summarized in Table 1. We include interactions between predictors motivated by our questions; significant interactions of interest were followed up using post-hoc tests using the *emmeans* package (Lenth 2024). For model selection, we first aimed to fit the model with the maximal random-effects structure (Barr et al., 2013; Bates et al., 2015a), given the data, in all cases. We then used the stepwise regression functionality of the *buildmer* package (Voeten, 2020) to trim this maximal model down to the optimal model based on likelihood ratio tests.

Table 1. Summary of predictors across experiments.

| predictor | summary | contrasts |
|---|---|---|
| step_n | The duration step of the segment | Numeric (centered) |
| original_length | The pre-manipulation segment length | short = -0.5, long = 0.5 |
| rate | The carrier speech rate | fast = -0.5, slow = 0.5 |
| segment_type | The target segment type | stop = -0.5, vowel = 0.5 |
| manipulation_type | The segments manipulated to achieve carrier rate changes | Exp. 2–4: all = -0.5, vowels = 0.5 Exp. 5: all = -0.5, consonants = 0.5 |
| subject_id | Unique participant label | random effect |
| word_pair | The target minimal pair | random effect |

## 3. Results

In Experiment 1, we aimed to confirm that perceptual rate normalization occurs for both stops and vowels in Japanese. We created a 9-step durational continuum for each base word (8 pairs * 2 words) and manipulated the duration of the carrier sentence to be 20% slower or 20% faster than the original production (2 rates), for 288 trials in total. If there is rate normalization, we would expect 'fast'

carriers to predict more 'long' responses, and we would expect 'slow' carriers to predict more 'short' responses.

The results of the linear mixed-effects regression model are displayed in Table 2, and the results are visualized in Figure 1. We find that rate plays a role, but that there is an interaction between rate and segment category, such that the effect of rate is weaker for vowels than for stops, even though vowels and consonants (including stops) are manipulated the same way in the carrier. Whether the stimuli were created from the original long or short word production also plays a large role, and this effect interacts with segment type, such that it is much stronger for vowels than for stops. This is likely due to pitch accent information: originally long vowel stimuli have a steeper pitch fall than originally short vowel stimuli (Kozasa, 2005). Post-hoc testing confirms that the effect of rate is significant across both vowels and stops and long and short baselines.

Table 2. Mixed-effects model output for Experiment 1. *Random effects kept by buildmer*: subject_id and word_pair, and by-subject and by-continuum random slopes for step_n, rate, and original_duration. *Nakagawa's R2*: conditional = 0.780, marginal = 0.683.

| | coef. | s.e. | p-value |
|---|---|---|---|
| (Intercept) | -0.06 | 0.22 | 0.8 |
| step_n | 1.05 | 0.04 | <0.01 |
| rate (slow-fast) | -2.03 | 0.08 | <0.01 |
| segment_type (vowel-stop) | -0.09 | 0.22 | 0.7 |
| original_length (long-short) | 2.54 | 0.21 | <0.01 |

| | | | |
|---|---|---|---|
| rate:seg_type | 0.41 | 0.15 | 0.01 |
| rate:original_length | 0.43 | 0.14 | <0.01 |
| seg_type:original_length | 1.97 | 0.4 | <0.01 |
| rate:seg_type:original_length | 0.07 | 0.27 | 0.81 |

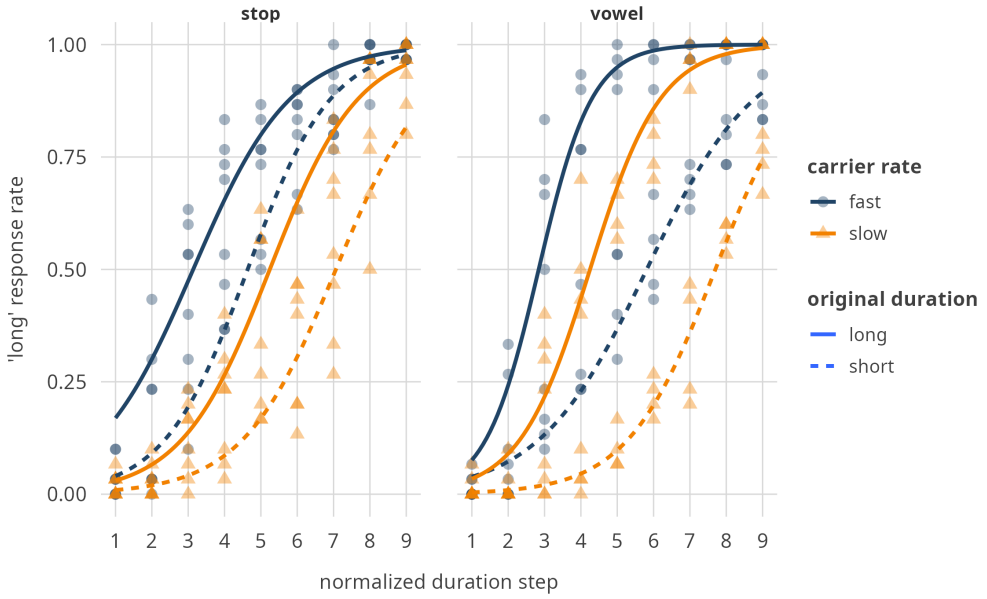



Fig. 1. The mean 'long' response rate for a given word pair at each duration rate[1], segment type, and original length for Experiment 1. The lines are the LOESS curves.

In Experiment 2, we investigated whether listeners attend to category-specific duration variation across rate conditions or use the global speech rate for rate normalization. To test this, we added a new condition in which the rate of the carrier sentence was increased or decreased by 20% by manipulating vowels only. This 20% overall change was achieved by changing the vowel duration by 35.5%, while keeping the duration of consonants the same. Each participant completed both

[1]Plots based on raw duration are also available in our OSF repository (see Data Availability Statement).

rate manipulation conditions. This doubled the number of trials, so to make the experiment more manageable, we reduced the number of steps in the durational continuum from nine to five, for 360 trials in total. If listeners normalize by attending to category-specific duration, we would expect the rate effect to differ depending on the manipulation condition and the segment type: perception of vowels would show a stronger effect when only vowels are altered than when all segments are manipulated (same as Experiment 1)., while stops would show a weaker effect for the former than the latter.

The results of the linear mixed-effects regression model are displayed in Table 3(a), and the results are visualized in Figure 2(a). We find that manipulation type interacts with rate and segment type. A post-hoc test comparing the fast to slow rate conditions shows that, as with Experiment 1, rate affects perception for both vowels and stops, and this effect holds across both manipulation types, suggesting that the rate effect can occur in the absence of category-specific duration change for stops. As in Experiment 1, original length also has a strong main effect, which is stronger for vowels than for consonants (not shown in Figure 2 (a) for readability).

Table 3. Mixed-effects model output for Experiments 2, 3 and 5. *Random effects kept by buildmer*: intercepts for subject_id and word_pair, and by-subject and by-continuum slopes for step_n, rate (E5 only), and original_duration (E2, E3 only).

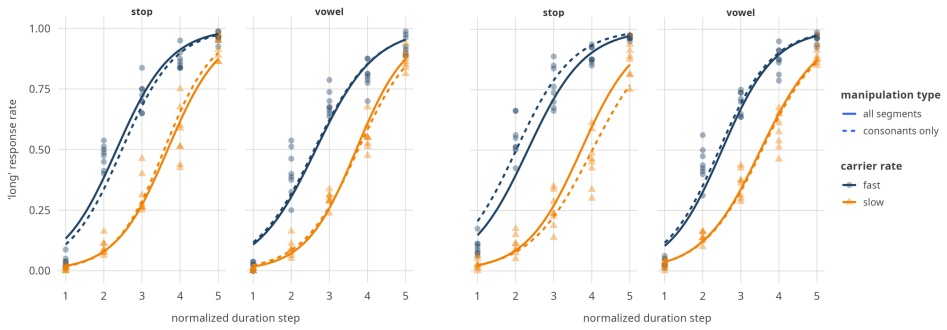
| | (a) Experiment 2 | | | (b) Experiment 3 | | | (c) Experiment 5 | | |
|---|---|---|---|---|---|---|---|---|---|
| Nakagawa's R2 | Cond. | | Marg. | Cond. | | Marg. | Cond. | | Marg. |
| | 0.791 | | 0.705 | 0.741 | | 0.633 | 0.785 | | 0.679 |
| | coef. | s.e. | p-value | coef. | s.e. | p-value | coef. | s.e. | p-value |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| (Intercept) | -0.15 | 0.19 | 0.41 | 3.83 | 0.38 | <0.01 | 3.8 | 0.32 | <0.01 |
| step_n | 2 | 0.08 | <0.01 | 1.8 | 0.09 | <0.01 | 1.9 | 0.1 | <0.01 |
| manipulation_type (subset-all) | -0.02 | 0.06 | 0.69 | -0.1 | 0.05 | 0.06 | 0.04 | 0.06 | 0.49 |
| rate (slow-fast) | -2.33 | 0.18 | <0.01 | -2.11 | 0.06 | <0.01 | -2.62 | 0.23 | <0.01 |
| segment_type (vowel-stop) | -0.63 | 0.16 | <0.01 | -0.74 | 0.21 | <0.01 | -0.45 | 0.2 | 0.03 |
| original_length (long-short) | 2.44 | 0.07 | <0.01 | 1.09 | 0.18 | <0.01 | 2.18 | 0.07 | <0.01 |
| manip:rate | 0.13 | 0.11 | 0.24 | 0.23 | 0.11 | 0.04 | -0.58 | 0.11 | <0.01 |
| manip:seg_type | 0.05 | 0.11 | 0.66 | 0.03 | 0.11 | 0.81 | 0.01 | 0.11 | 0.94 |
| rate:seg_type | 0.05 | 0.29 | 0.87 | 0 | 0.12 | 0.99 | 0.76 | 0.48 | 0.11 |
| seg_type:orig_leng | 1.41 | 0.13 | <0.01 | -0.06 | 0.37 | 0.87 | 1.41 | 0.13 | <0.01 |
| manip:rate:seg_type | -0.55 | 0.23 | 0.02 | -0.57 | 0.22 | 0.01 | 0.73 | 0.22 | <0.01 |

Fig. 2. The mean 'long' response rate for a given word pair at each duration step[2] by rate, segment type, and manipulation type for (a) Experiment 2 and (b) Experiment 5. The lines are the LOESS curves.

[2]Plots based on raw duration are also available in our OSF repository (see Data Availability Statement).

(a) (b)




A second post-hoc test comparing the effect across segment types and manipulation types shows that rate has a weaker effect for stops in the vowel-only manipulation condition compared to the overall manipulation condition. In Figure 2(a), for stops, the two solid lines (overall manipulation) are further apart than the two dotted lines (vowel-only manipulation). However, the rate effect on vowels does not vary significantly by manipulation type: the two solid lines are no less separated from each other than the two dotted lines. If rate normalization is category-dependent, why do stops show sensitivity to manipulation type but not vowels?

One hypothesis is that because pitch accent is a strong cue to vowel length, it may play a larger role for vowels, making them less sensitive to speech-rate variation in general, so that the finer distinction between the two manipulation conditions does not affect perception. We test this hypothesis in Experiment 3, which replicates Experiment 2, but with all pitch information removed by flattening the pitch of all stimuli. Since the pitch cue is no longer available, we expect that we should find a difference between the vowel-only and overall rate manipulation conditions for vowels. The results of the linear mixed-effects regression model are displayed in

Table 3(b). We no longer find an interaction of original length with segment type. That is, vowels are no more affected by original length than stops, indicating that the loss of pitch cue for length in vowels makes vowels and stops more comparable in this regard. Otherwise, however, the results are essentially the same as in Experiment 2. There is still a difference between manipulation types for stops, and no difference between manipulation types for vowels. In other words, pitch information seems to explain the stronger effect of original length for vowels but cannot explain the lack of a manipulation-type effect for vowels.

Another possibility is that, because the two manipulation conditions were intermingled and presented in pseudorandom order, participants did not cue in to the difference between the conditions. We test this hypothesis in Experiment 4 (not shown), which replicates Experiment 3 with a blocked design. Yet again, we find the results are largely the same as Experiments 2 and 3; whether the stimuli were presented in a blocked or unblocked order did not make a difference.

One final possibility is that the two manipulation types are not actually equally different with respect to vowels and consonants. For consonants, the difference is qualitative: there is no change in consonant duration in the vowel-only manipulation, versus there being some change in consonant duration in the global manipulation. For vowels, though, the difference is quantitative: there are changes in vowel duration in both manipulation conditions, with the difference just being between smaller and larger changes. It is perhaps not surprising, then, that consonants are more noticeably affected by the change in manipulation methods, while vowels are not. Experiment 5 followed up on this by inverting the way consonants and vowels are manipulated: rather than one condition where only

vowels are manipulated and another where both segment types are manipulated, we had one condition where consonants are manipulated and another where both segment types are manipulated. If listeners normalize using category-specific duration information, we would expect perception of vowels, not stops, to show different rate effects by manipulation type. The results of the linear mixed-effects regression model are displayed in Table 3(c), and the results are visualized in Figure 2(b). Post-hoc testing finds a significant rate effect for both vowels and stops in both manipulation conditions, which means that vowels are affected by rate even when only consonants are altered between the slow and fast rates, similar to the rate effect for stops in the vowel-only manipulation condition in Experiments 2–4. The result of main interest is an interaction between rate, manipulation type and segment type, which is significant. Post hoc testing shows that this is due to a difference in responses to stop segment trials: the effect of rate for stops is greater when only consonants are manipulated than when all segments are manipulated, while the rate effect for vowels remains the same. In other words, the different effects of manipulation for vowels and stops found in Experiments 2–4 persisted regardless of whether the manipulation was varied quantitatively (more or less duration change between speech rates) or qualitatively (change or no change).

## 4. Discussion

This study has demonstrated that perception of both vowels and stops is affected by rate normalization in Japanese and found a robust rate effect across all experiments and all conditions, even when rates were varied by altering vowels or consonants only, leaving the other segment types unchanged. For vowels, we only

found the effect of the overall speech rate and no effect of category-specific duration variation. These findings are consistent with the neural entrainment account of rate effect (e.g., Kösem et al., 2018), whereby the neural oscillation tracks the overall rhythmic fluctuation of the speech signal and creates predictions for the upcoming speech. However, we also found a small but significant effect of manipulation type, but only for stops. The effect of rate normalization is weaker for stops when only vowel durations are manipulated (Experiments 2–4); and it is stronger when stop durations are manipulated to a higher extent (Experiment 5). This suggests that one of the cues listeners attend to when perceiving the length of a stop is the duration of other stops in that category in the speech stream, and this cue-specific duration variation can modulate the overall speech rate effect. This cue-specific effect is consistent with episodic accounts of rate normalization, which hold that listeners attend not (just) to overall rate, but to the duration of specific segments.

What remains puzzling is the asymmetry between vowels and stops. One possibility is the different ways that vowel and consonant gestures are organized in speech production. A body of articulatory phonetic studies has established that vocalic gestures are articulated contiguously through intervening consonantal gestures, while consonantal gestures are superimposed on top of a sequence of continuous vocalic gestures (Öhman, 1966; Carney and Moll, 1971; Fowler, 1983; Browman and Goldstein, 1990; Gafos, 1996). Under this conception of speech production, even when speech rate is modulated by stretching the duration of consonants only, the speaker-turned-listener reconstructs the vocalic gestures present underneath the superimposed consonantal gestures, and thus perceives the lengthening and shortening of vocalic gestures along with the consonantal ones.

Another possibility is that listeners start off with different expectations about the duration of the target vowels and consonants — and how they would vary by speech rate — based on their lifetimes of exposure, and the speech material in the carrier sentence can only modulate this expectation, not completely change it. If so, then what prior belief would listeners have to hold to exhibit the vowel-stop asymmetry observed in our study? Some have claimed that vowels generally vary more by speech rate than consonants in production in Japanese (Kuwabara, 1996) and other languages (Gay, 1978; Lo and Sóskuthy, 2023). If this is the case, the listeners may have more confidence about adjusting their vowel perception when they recognize fast or speech rate compared to stops and for stops, they may rely less on speech rate per se of the ambient speech, but the absolute duration of stops is relatively more important compared to vowels. The stronger rate effect for stops than vowels found in Experiment 1 may seem to contradict this prediction, but note that our rate manipulation, which varies the duration of vowels and consonants equally, may artificially exaggerate the consonant duration difference by rate, if they vary less than vowels in naturalistic fast and slow speech.

Indeed, rate normalization is not necessary or consistently attested in all durationally-signaled contrasts. Nakai and Scobbie (2016) found that the category boundary for the English voicing distinction does not meaningfully vary by speech rate and that there is no overlap between the VOT of voiced and voiceless, obviating the need for rate normalization. Theodore et al. (2009) found that speakers varied widely in the magnitude of VOT change by speech rate, with some exhibiting a large difference by speech rate while others showing almost no difference; this cross-speaker variation makes rate normalization less reliable. Relatedly, Ting and Kang

(2023) found that after listening to a short dialogue between two speakers, listeners did not appear to use each speaker's habitual speech rate as a cue for distinguishing /pi/ and /bi/, unlike German /a/ and /a:/ (Reinisch 2016). Future studies need to investigate the rate effects in the production of specific segment types in specific languages, and how (or if) their degree of variability in production informs their rate normalization in perception.


## Acknowledgments

[to be added in the event of acceptance]


## Author Declarations

### *Conflict of Interest*

The authors have no conflicts to disclose.

### *Ethics Approval*

The human-subjects component of the research was approved by the University of Toronto Office of Research Ethics. Informed consent was obtained from all participants.

## Data Availability

The data that support the findings of this study, along with the experimental materials, are openly available in an Open Science Framework (osf.io) repository at https://doi.org/10.17605/OSF.IO/SB8UF, reference number SB8UF.

## References and Links


Amano, S., and Hirata, Y. (**2010**). "Perception and production boundaries between single and geminate stops in Japanese," The Journal of the Acoustical Society of America, **128**, 2049–2058.

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (**2013**). "Random effects structure for confirmatory hypothesis testing: Keep it maximal," Journal of memory and language, **68**, 255–278.

Bates, D., Kliegl, R., Vasishth, S., and Baayen, H. (**2015a**). "Parsimonious mixed models," arXiv preprint arXiv:1506.04967.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (**2015b**). "Fitting linear mixed-effects models using lme4," Journal of Statistical Software, **67**, 1–48. doi:10.18637/jss.v067.i01

Browman, C. P., and Goldstein, L. (**1990**). "Tiers in articulatory phonology, with some implications for casual speech," Papers in laboratory phonology I: Between the grammar and physics of speech, **1**, 341–397.

Carney, P. J., and Moll, K. L. (**1971**). "A cinefluorographic investigation of fricative consonant-vowel coarticulation," Phonetica, **23**, 193–202.

de Leeuw, J. R. (**2015**). "jsPsych: A JavaScript library for creating behavioral experiments in a Web browser," Behavior research methods, **47**, 1–12.

Fowler, C. A. (**1983**). "Converging sources of evidence on spoken and perceived rhythms of speech: Cyclic production of vowels in monosyllabic stress feet." Journal of Experimental Psychology: General, **112**, 386.

Gafos, A. I. (**1996**). "V/C planar segregation and long-distance spreading revisited," Selected papers on theoretical and applied linguistics, **9**, 25–37.

Gay, T. (**1978**). "Effect of speaking rate on vowel formant movements," The Journal of the Acoustical Society of America, **63**, 223–230.

Hawkins, O. (**2024**). *Pilot: A minimal ggplot2 theme with an accessible discrete color palette*. R package version 4.0.1. Retrieved from https://github.com/olihawkins/pilot

Heffner, C. C., Newman, R. S., and Idsardi, W. J. (**2017**). "Support for context effects on segmentation and segments depends on the context," Attention, Perception, & Psychophysics, **79**, 964–988.

Hirata, Y. (**2004**). "Effects of speaking rate on the vowel length distinction in japanese," Journal of Phonetics, **32**, 565–589.

Hirata, Y., and Lambacher, S. G. (**2004**). "Role of word-external contexts in native speakers' identification of vowel length in Japanese," Phonetica, **61**, 177–200.

Hirata, Y., and Whiton, J. (**2005**). "Effects of speaking rate on the single/geminate stop distinction in Japanese," The Journal of the Acoustical Society of America, **118**, 1647–1660.

Idemaru, K., and Guion-Anderson, S. (**2010**). "Relational timing in the production and perception of Japanese singleton and geminate stops," Phonetica, **67**, 25–46.

Kawahara, S., Kato, M., and Idemaru, K. (**2022**). "Speaking rate normalization across different talkers in the perception of Japanese stop and vowel length contrasts," JASA Express Letters, **2**.

Kleinschmidt, D. F., and Jaeger, T. F. (**2015**). "Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel," Psychological Review, **122**, 148.

Kösem, A., Bosker, H. R., Takashima, A., Meyer, A., Jensen, O., and Hagoort, P. (**2018**). "Neural entrainment determines the words we hear," Current Biology, **28**, 2867–2875.

Kozasa, T. (**2005**). "An acoustic and perceptual investigation of long vowels in Japanese and Pohnpeian," Ph.D. dissertation, University of Hawai'i at Manoa, Honolulu, Hawai'i.

Kuwabara, H. (**1996**). "Acoustic properties of phonemes in continuous speech for different speaking rate," Proceeding of fourth international conference on spoken language processing. ICSLP'96, IEEE, 2435–2438.

Lahiri, A., Gewirth, L., and Blumstein, S. E. (**1984**). "A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants: Evidence from a cross-language study," The Journal of the Acoustical Society of America, **76**, 391–404.

Lenth, R. V. (**2024**). *Emmeans: Estimated marginal means, aka least-squares means*. R packaRetrieved from https://CRAN.R-project.org/package=emmeans

Lo, R. Y.-H., and Sóskuthy, M. (**2023**). "Articulation rate in consonants and vowels: Results and methodological challenges from a cross-linguistic corpus study," In R. Skarnitzl and J. Volín (Eds.), Proceedings of the 20th international congress of phonetic sciences, Guarant International, 3206–3210.

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

Nakai, S., and 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**, 1–31.

Öhman, S. E. (**1966**). "Coarticulation in VCV utterances: Spectrographic measurements," The Journal of the Acoustical Society of America, **39**, 151–168.

Pierrehumbert, J. (**2001**). "Lenition and contrast," Frequency and the emergence of linguistic structure, **45**, 137.

Port, R. F. (**1981**). "Linguistic timing factors in combination," The Journal of the Acoustical Society of America, **69**, 262–274.

Port, R. F., Al-Ani, S., and Maeda, S. (**1980**). "Temporal compensation and universal phonetics," Phonetica, **37**, 235–252.

R Core Team (**2021**). *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/

Reinisch, E., and Sjerps, M. J. (**2013**). "The uptake of spectral and temporal cues in vowel perception is rapidly influenced by context," Journal of Phonetics, **41**, 101–116.

Theodore, R. M., Miller, J. L., and DeSteno, D. (**2009**). "Individual talker differences in voice-onset-time: Contextual influences," The Journal of the Acoustical Society of America, **125**, 3974–3982.

Ting, C., and Kang, Y. (**2023**). "The effect of habitual speech rate on speaker-specific processing in english stop voicing perception," Language and Speech.

Voeten, C. C. (**2020**). "Buildmer: Stepwise elimination and term reordering for mixed-effects regression," R package version 2.11. Retrieved from https://CRAN.R-project.org/package=buildmer

Wickham, H. (**2016**). *ggplot2: Elegant graphics for data analysis*, Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org