JASA EXPRESS LETTERS ARTICLE asa.scitation.org/journal/jel

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

Shigeto Kawahara,[1] **Misaki Kato,**[2] **and Kaori Idemaru**[3]
[1]*Institute of Cultural and Linguistic Studies, 2-15-45 Mita Minato-ku Tokyo, 108-8345, Japan*
[2]*Department of Linguistics, 1290 University of Oregon, Straub Hall, Eugene, Oregon 97403, USA*
[3]*Department of East Asian Languages and Literatures, 1248 University of Oregon, Friendly Hall, Eugene, Oregon 97403, USA*
kawahara@icl.keio.ac.jp, misaki@uoregon.edu, 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 is talker-specific is understudied. This experiment investigated whether Japanese 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: a singleton-geminate stop contrast and short-long vowel contrast in Japanese. The results suggest that listeners generalized their rate-based adjustments to different talkers' speech regardless of whether the target contrasts depended on silent closure duration or vowel duration. Our results thus support the view that speaking rate normalization is an obligatory process that happens in the early phase of perception.

[Editor: Martin Cooke] https://doi.org/10.1121/10.0009793

Received: 10 November 2021 **Accepted:** 24 February 2022 **Published Online:** 16 March 2022

1. Introduction

Even within a single language, there are significant variations in how people speak. Speed is one aspect of speech that varies considerably; some people speak faster than others (Crystal and House, 1988; Quené, 2008), and the same person may speak faster or slower on different occasions (Miller *et al.*, 1984). It has been demonstrated that listeners take this variation in speaking rate into account when processing speech (Bosker, 2017; Hirata and Whiton, 2005; Idemaru and Guion-Anderson, 2010; Reinisch *et al.*, 2011a; Summerfield, 1981; Wayland *et al.*, 1994). One piece of evidence for rate-dependent speech perception is 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 rate of the surrounding speech (Kidd, 1989; Miller and Dexter, 1988; Sawusch and Newman, 2000). 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 precursor phrase than a slower precursor phrase (Gordon, 1988). Similar effects have been found in perception of other contrasts involving temporal cues (Reinisch, 2016; Reinisch and Sjerps, 2013), manner of articulation (Wade and Holt, 2005), lexical stress (Reinisch *et al.*, 2011a), word segmentation (Reinisch *et al.*, 2011b), as well as in the perception of function words (Baese-Berk *et al.*, 2014; Dilley and Pitt, 2010).

One question that arises is how strongly listeners' rate-based normalization is associated with a specific talker's voice. That is, are listeners' auditory normalizations of phonetic temporal cues based on general auditory input (e.g., speech produced by multiple talkers) or are they based on the speech produced by a specific talker? Newman and Sawusch (2009) provide considerable insight into this question. By testing rate-dependent perception of the English /k/-/g/ continuum, the authors have demonstrated that English listeners use speech rate information pertaining to one voice (e.g., the precursor in a male voice) to adjust their perception of another voice (e.g., the target in a female voice). Such cross-talker rate-normalization was observed even when the precursor and the target in different voices were delivered to listeners' different ears, indicating that different spatial location of the talkers did not block the occurrence of cross-talker rate-normalization. Rate normalization of the irrelevant voice was blocked, albeit not completely, only when rate information was simultaneously available in both the irrelevant and the relevant voices. Given these results, and given the results of their previous work and that of others (Miller and Dexter, 1988; Newman and Sawusch, 1996; Sawusch and Newman, 2000), the authors have proposed that speaking rate normalization is an obligatory process that happens in the early phase of perception, arguably before talker-specific information is processed.

If this conclusion by Newman and Sawusch is on the right track, rate normalization across different talkers should occur with other duration-cued contrasts and in other languages. To test this prediction, the current study extends

the scope of the hypothesis of Newman and Sawusch (2009) by examining the perception of two duration-cued contrasts in Japanese. Japanese has a singleton-geminate consonant contrast (e.g., /k/-/kk/) as well as a short-long vowel contrast (e.g., /e/-/ee/). Both of these contrasts are primarily based on durational differences (Vance, 2008) and rate-dependent perception has been reported at least for the singleton-geminate contrast (Hirata and Whiton, 2005; Idemaru and Guion-Anderson, 2010). Whereas the English /k/-/g/ continuum in Newman and Sawusch (2009) varied from 14 to 72 ms along the VOT continuum (with a difference of 58 ms), the critical duration of the Japanese stop and vowel length contrasts may vary in the range of 80–100 ms (e.g., the duration varies from 60 ms to 140 ms in the current study) [see also Kawahara (2015) for a summary of previous acoustic studies]. While the typical rate effects were around 3%–4% in Newman and Sawusch (2009), the larger variation in the critical feature in Japanese may provide an opportunity to observe more robust rate normalization effects across talkers.

Furthermore, the two contrasts examined here (i.e., singleton-geminate stops and short-long vowels) differ in terms of the content of the duration information. While the singleton-geminate contrast is cued by the duration of the silent closure, the short-long vowel contrast is cued by the duration of the intervals that are "spectrally rich," which may include talker-specific information. This subtle difference might result in a different degree of rate-based adjustment [cf. Kraljic and Samuel (2005), (2006), (2007)]. However, given the consistent cross-talker rate normalization found in Newman and Sawusch (2009), and also given the fact that the target stop is surrounded by vowels in our test word, the likelihood that we may find differences between the stop and the vowel target may be slim. Nonetheless, the basic and the potentially critical difference between the two target segments, namely, that one contains while the other lacks talker information, is an interesting possibility to explore.

2. Methods

2.1 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. None of them reported a history of speech or hearing impairment.

2.2 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: /heko-hekko/ (consonant) and /hesu-heesu/ (vowel). Two native Japanese talkers (1 female, 1 male), who did not participate in the experiment as listeners, 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 Tokyo 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 (Boersma and Weenink, 2015). 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: e.g., Kubozono (2006)]. 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. The precursor and target word durations of the two selected talkers were adjusted to be the mean durations of their productions (the "normal" speech rate condition). 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 = 0.625) with PSOLA, resulting in three rates: fast, normal, 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 encompassed typical short and long segments (Kawahara, 2015). 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 × 2 talkers) were combined with all target words (2 segments × 5 durations × 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.

2.3 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 (Peirce, 2007). 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 was 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 min.

2.4 Analysis

A mixed effects Bayesian logistic regression model was fit to the data. The dependent variable was the binary long vs short responses. The fixed factors were (1) centered duration, (2) segment type (vowel vs consonants), (3) precursor speed (fast vs normal; fast vs slow), and (4) the precursor congruency as well as all their interaction terms. A random intercept for participants as well as a random slope for all the fixed factors and their interaction terms were included in the model, as Bayesian analyses are less likely to face convergence issues than corresponding frequentist analyses [e.g., Eager and Joseph (2017)], especially when regularizing priors were used.

The prior for the intercept was set to be normal (0, 1) weakly informative priors (Lemoine, 2019). For all fixed effect coefficients, we use a Cauchy prior with scale = 2.5 (Gelman *et al.*, 2008). The analysis was implemented using the BRMS package (Bürkner, 2017) and R (ver. 4.0.2). Four chains were run, each with 2000 iterations, the first half of which were discarded as warmups. All the R-hat values associated with the fixed effects were 1.00, suggesting that the four chains mixed successfully. The raw data, the analysis script, the whole regression result as well as all the posterior samples of the estimates are available at OSF (2022). In this paper, we report 95% Credible Intervals (95% CrI) for the effects that are of interest as well as those effects that were found to be credible, and refer the readers to the markdown file that is available at the OSF repository for the full detail. While Bayesian analyses do not force us to interpret the results in terms of a "significant vs non-significant" dichotomy, we interpret a particular effect to be meaningful as long as its 95% CrI does not contain 0.

One advantage of resorting to a Bayesian analysis is that it allows us to access how credible a null effect is, whereas in frequentist analyses, the best we can conclude is that we cannot reject the null hypothesis (Gallistel, 2009). To this end, we deployed a ROPE (region of practical equivalence) analysis (Kruschke and Liddell, 2018). In this approach, we define a region that is practically equivalent to a point hypothesis (here $\beta = 0$). For the current purpose we took the width of this region to be a negligible standardized effect size of 0.1 in the sense of Cohen (1988), which is 0.18 in logistic regression analyses (Makowski *et al.*, 2019). In short, if a posterior distribution of an estimate of a particular parameter is fully contained in [−0.18, 0.18], we accept that null effect (Kruschke and Liddell, 2018). The markdown file also reports, for each estimated parameter, how many posterior samples are contained in this ROPE.

3. Results

Figure 1 illustrates the proportion of long responses (i.e., /hekko/ or /heesu/) across the five steps of the target segment duration continua and three levels of precursor rate (fast, normal, or slow) by congruency (congruent vs incongruent). We first present the results involving those of consonant, which was the reference level for the factor segment. As expected, the effect of duration was credible in such a way that longer duration induced more "long" responses ($\beta = 0.12$, 95% CrI [0.09, 0.14]). The difference between the fast precursor and normal precursor was also meaningful ($\beta = -1.24$, 95% CrI [−1.60, −0.89]), with the latter inducing less "long" responses. The difference between the fast precursor and slow precursor was also credible ($\beta = -1.22$, 95% CrI [−1.64, −0.82]) in the same direction. We thus observed robust normalization effects in our results in that the perception of target duration depended on the speaking rate of the precursor.

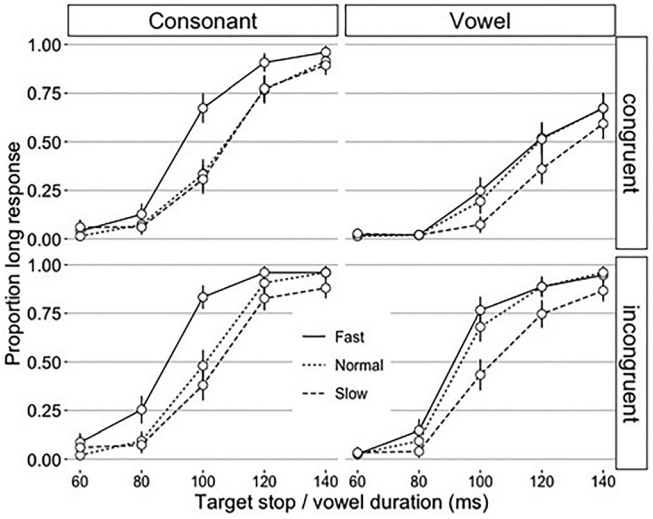

Fig. 1. Proportion of "long" responses for two segments (consonant /hekko/ or vowel /heesu/) for three levels of precursor rate (fast, normal, slow) across five steps of target stop closure or vowel duration (ms) by congruency (congruent vs incongruent). Error bars indicate the 95% confidence interval of the mean.

The estimate for interaction between the consonant duration and the congruency ($\beta = -0.01$, 95% CrI [−0.03, 0.02]) was fully contained in the ROPE of [−0.18, 0.18], which suggests that the way the duration continuum was perceived did not differ between the congruent condition and incongruent condition. More relevant to our research question, the two three-way interactions between duration × congruency × fast vs normal ($\beta = 0.02$, 95% CrI [−0.01, 0.05]) and duration × congruency × fast vs slow ($\beta = 0.00$, 95% CrI [−0.02, 0.03]) can also considered to be null, suggesting that the degrees to which the three precursors affected the perception of consonant duration did not differ between the congruent and incongruent conditions. In other words, listeners adjusted their perception of duration depending on the rate of the precursor to the same degree whether the target and the precursor were spoken in the same voice or different voices. This result is consistent with the hypothesis advanced by Newman and Sawusch (2009).

Furthermore, the two four-way interactions, duration × congruency × fast vs normal × consonant vs vowel ($\beta = -0.02$, 95% CrI [−0.05, 0.02]) as well as duration × congruency × fast vs slow × consonant vs vowel ($\beta = -0.02$, 95% CrI [−0.06, 0.01]) can also considered to be both null. These results demonstrate that the effects of congruency on rate-based normalization did not differ between stop-based contrasts and vowel-based contrasts. Here, we extend the scope of the conclusion by Newman and Sawusch (2009) in that cross-talker adjustment of the target duration is observed whether the target is a stop or a vowel. The talker-specific information contained in the vowel, to the extent that it is present in our stimuli, does not block or modulate the cross-talker normalization of segmental duration [though see also Kraljic and Samuel (2005), (2006), (2007)].

Albeit being less directly relevant to our research question, we have found some complicated patterns in our results as well. Congruency had a credible effect in such a way that incongruent precursors induced more "long" responses ($\beta = 0.80$, 95% CrI [0.28, 1.36]), for which we do not have a clear explanation. Vowels were generally less likely to be judged as "long" than consonants ($\beta = -2.06$, 95% CrI [−2.41, −1.71]). The effects of duration meaningfully interacted with consonant vs vowel distinction such that the effects of duration increase were less pronounced in vowels ($\beta = -0.04$, 95% CrI [−0.06, −0.02]). The effect of congruency interacted with the consonant vs vowel distinction such that the congruency difference was more pronounced in vowels than in consonants ($\beta = 1.39$, 95% CrI [0.92, 1.89]). The difference between the effect of fast vs normal also differed between vowels and consonants ($\beta = 1.02$, 95% CrI [0.52, 1.51]). Whereas we have no explanation for the varying effect of congruency on the response, those involving differences across the consonant vs vowel may be due to the design of the stimuli. While we used the range varying from 60 to 140 ms for both vowel and stop duration, this range may not optimally fit the natural variation of vowel duration. That may have caused fewer "long" responses to vowel continua and other effects involving vowel. A follow-up experiment using a different duration range is necessary to address this possibility.

4. Discussion

The present study investigated whether listeners' perception of temporally contrastive phonemes is influenced by the speaking rate of the precursor phrase when the speaker of the precursor and the target word match (congruent) and mismatch (incongruent), and whether this pattern differs for different target contrasts: the Japanese singleton-geminate stop contrast (i.e., /k/-/kk/) and the short-long vowel contrast (i.e., /e/-/ee/). The results demonstrated the effect of precursor rates in both congruent and incongruent conditions, and this pattern persisted for both contrasts. That is, the faster the precursor rate, the more often the target phoneme was perceived as the "long" phoneme (i.e., geminate stop /kk/ and long vowel /ee/) even when the speaker of the precursor phrase differed from the speaker of the target word.

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. These results are consistent with the results of Newman and Sawusch (2009), and in turn seem to support the proposal that rate normalization is an obligatory process, where listeners use any available information to make rate-based adjustments (Miller and Dexter, 1988; Newman and Sawusch, 2009; Sawusch and Newman, 2000). Our current study succeeded in extending the scope of this proposal by using data from a language other than English and with novel contrast types. We also found larger rate normalization effects (i.e., the difference in the percentage of long response between the fast and slow rate) for the Japanese short-long contrasts (15% for the consonant; 11% for the vowel) than were reported for the English /k/-/g/ contrast [3%–4%, Newman and Sawusch (2009)].

Studies have shown that listeners' rate-based adjustments occur even for the irrelevant talker's voice when presented simultaneously with the relevant talker's voice (Newman and Sawusch, 2009) and under conditions with varying attentional demands (Bosker *et al.*, 2017). The present results contribute to these lines of research demonstrating that rate normalization across talkers occurs with the target segment that 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 (Bosker *et al.*, 2017; Kingston *et al.*, 2009; Sawusch and Newman, 2000; Wade and Holt, 2005) That is, extraction of rate information may occur earlier than segregation of voices, with the rate information affecting subsequent auditory processing.

Given this seemingly obligatory nature of rate-based speech perception, future research may address what factors block cross-talker rate normalization. In other words, when listeners are faced with situations in which multiple people are

speaking, constantly adjusting their perception strategies regardless of the source of the voice does not seem to be an efficient process. Newman and Sawusch (2009) have shown that rate information in the relevant voice that is simultaneously available with the rate information in an irrelevant voice reduces the use of the irrelevant voice. Prior training or habituation to the relevant voice, visual display of the relevant speaker (if gender distinguishes the talkers), or disruption of the speech flow (e.g., a long pause between the precursor and the target) may block or modulate the application of cross-talker rate normalization.

Acknowledgments

We thank Miko Suzuki and Hayli Brown for their help running research participants. We also thank the audience of Speech and Prosody 2020 and two anonymous reviewers for their helpful comments. The research was partially supported by JSPS Grant No. 20H05617.

References and links

Baese-Berk, M. M., Heffner, C. C., Dilley, L. C., Pitt, M. A., Morrill, T. H., and McAuley, J. D. (**2014**). "Long-term temporal tracking of speech rate affects spoken-word recognition," Psych. Sci. **25**(8), 1546–1553.

Boersma, P., and Weenink, D. (**2015**). "Praat: Doing phonetics by computer [computer program]," http://www.praat.org.

Bosker, H. R. (**2017**). "Accounting for rate-dependent category boundary shifts in speech perception," Atten. Percept. Psychophys. **79**(1), 333–343.

Bosker, H. R., Reinisch, E., and Sjerps, M. J. (**2017**). "Cognitive load makes speech sound fast, but does not modulate acoustic context effects," J. Mem. Lang. **94**, 166–176.

Bürkner, P.-C. (**2017**). *brms: An R Package for Bayesian Multilevel Models* Using *Stan*, R package.

Cohen, J. (**1988**). *Statistical Power Analysis for the Behavioral Science* (Lawrence Erlbaum, Mahwah, NJ).

Crystal, T. H., and House, A. S. (**1988**). "Segmental durations in connected-speech signals: Current results," J. Acoust. Soc. Am. **83**(4), 1553–1573.

Dilley, L. C., and Pitt, M. A. (**2010**). "Altering context speech rate can cause words to appear or disappear," Psych. Sci. **21**(11), 1664–1670.

Eager, C., and Joseph, R. (**2017**). "Mixed effects models are sometimes terrible," MS thesis, University of Illinois at Urbana Champaign, Champaign, IL.

Gallistel, R. C. (**2009**). "The importance of proving the null," Psychol. Rev. **116**(2), 439–453.

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. (**2008**). "A weakly informative default prior distribution for logistic and other regression models," Ann. Appl. Stat. **2**, 1360–1383.

Gordon, P. C. (**1988**). "Induction of rate-dependent processing by coarse-grained aspects of speech," Perc. Psychophys. **43**(2), 137–146.

Hirata, Y., and Whiton, J. (**2005**). "Effects of speaking rate on the single/geminate stop distinction in Japanese," J. Acoust. Soc. Am. **118**(3), 1647–1660.

Idemaru, K., and Guion-Anderson, S. (**2010**). "Relational timing in the production and perception of Japanese singleton and geminate stops," Phonetica **67**(1-2), 25–46.

Kawahara, S. (**2015**). "The phonetics of *sokuon*, obstruent geminates," in *The Handbook of Japanese Language and Linguistics: Phonetics and Phonology*, edited by H. Kubozono (Mouton, Berlin, Germany), pp. 43-73.

Kidd, G. R. (**1989**). "Articulatory-rate context effects in phoneme identification," J. Exp. Psych.: Human Percept. Perform. **15**(4), 736–748.

Kingston, J., Kawahara, S., Chambless, D., Mash, D., and Brenner-Alsop, E. (**2009**). "Contextual effects on the perception of duration," J. Phon. **37**(3), 297–320.

Kraljic, T., and Samuel, A. G. (**2005**). "Perceptual learning for speech: Is there a return to normal?," Cog. Psych. **51**(2), 141–178.

Kraljic, T., and Samuel, A. G. (**2006**). "Generalization in perceptual learning for speech," Psychon. Bull. Rev. **13**(2), 262–268.

Kraljic, T., and Samuel, A. G. (**2007**). "Perceptual adjustments to multiple speakers," J. Mem. Lang. **56**(1), 1–15.

Kruschke, J. K., and Liddell, T. M. (**2018**). "The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective," Psychol. Bull. Rev. **25**, 178–206.

Kubozono, H. (**2006**). "Where does loanword prosody come from?: A case study of Japanese loanword accent," Lingua **116**, 1140–1170.

Lemoine, N. P. (**2019**). "Moving beyond noninformative priors: Why and how to choose weakly informative priors in Bayesian analyses," Oikos **128**, 912–928.

Makowski, D., Ben-Shachar, M. S., and Lüdecke, D. (**2019**). "bayestestr: Describing effects and their uncertainty, existence and significance within the Bayesian framework," J. Open Source Softw. **4**(40), 1–8.

Miller, J. L., and Dexter, E. R. (**1988**). "Effects of speaking rate and lexical status on phonetic perception," J. Exp. Psych.: Human Percept. Perform. **14**(3), 369–378.

Miller, J. L., Grosjean, F., and Lomanto, C. (**1984**). "Articulation rate and its variability in spontaneous speech: A reanalysis and some implications," Phonetica **41**(4), 215–225.

Newman, R. S., and Sawusch, J. R. (**1996**). "Perceptual normalization for speaking rate: Effects of temporal distance," Percept. Psychophys. **58**(4), 540–560.

Newman, R. S., and Sawusch, J. R. (**2009**). "Perceptual normalization for speaking rate III: Effects of the rate of one voice on perception of another," J. Phon. **37**(1), 46–65.

OSF (**2022**). https://osf.io/vb7et/ (Last viewed March 9, 2022).

Peirce, J. W. (**2007**). "PsychoPy-Psychophysics software in Python," J. Neuro. Methods **162**(1–2), 8–13.

Quené, H. (**2008**). "Multilevel modeling of between-speaker and within-speaker variation in spontaneous speech tempo," J. Acoust. Soc. Am. **123**(2), 1104–1113.

Reinisch, E., Jesse, A., and McQueen, J. M. (**2011a**). "Speaking rate from proximal and distal contexts is used during word segmentation," J. Exp. Psych.: Human Percept. Perform. **37**(3), 978–996.
Reinisch, E., Jesse, A., and McQueen, J. M. (**2011b**). "Speaking rate affects the perception of duration as a suprasegmental lexical-stress cue," Lang. Speech **54**(2), 147–165.
Reinisch, E., and Sjerps, M. J. (**2013**). "The uptake of spectral and temporal cues in vowel perception is rapidly influenced by context," J. Phon. **41**(2), 101–116.
Reinisch, E. V. A. (**2016**). "Speaker-specific processing and local context information: The case of speaking rate," Appl. Psycholing. **37**(6), 1397–1415.
Sawusch, J. R., and Newman, R. S. (**2000**). "Perceptual normalization for speaking rate II: Effects of signal discontinuities," Perc. Psychophys. **62**(2), 285–300.
Summerfield, Q. (**1981**). "Articulatory rate and perceptual constancy in phonetic perception," J. Exp. Psychol.: Human Percept. Perform. **7**(5), 1074–1095.
Vance, T. J. (**2008**). *The Sounds of Japanese with Audio CD* (Cambridge University Press, Cambridge).
Wade, T., and Holt, L. L. (**2005**). "Perceptual effects of preceding nonspeech rate on temporal properties of speech categories," Perc. Psychophys. **67**(6), 939–950.
Wayland, S. C., Miller, J. L., and Volaitis, L. E. (**1994**). "The influence of sentential speaking rate on the internal structure of phonetic categories," J. Acoust. Soc. Am. **95**(5), 2694–2701.