SYLLABLE RATE DRIVES RATE NORMALIZATION, BUT IS NOT THE ONLY FACTOR

Giulio G.A. Severijnen [1], Hans Rutger Bosker [1,2], James M. McQueen [1,2]

[1]Donders Institute for Brain, Cognition, and Behaviour, Radboud University Nijmegen, The Netherlands
[2]Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
giulio.severijnen@donders.ru.nl, hansrutger.bosker@donders.ru.nl, james.mcqueen@donders.ru.nl

ABSTRACT

Speech is perceived relative to the speech rate in the context. It is unclear, however, what information listeners use to compute speech rate. The present study examines whether listeners use the number of syllables per unit time (i.e., syllable rate) as a measure of speech rate, as indexed by subsequent vowel perception. We ran two rate-normalization experiments in which participants heard duration-matched word lists that contained either monosyllabic vs. bisyllabic words (Experiment 1), or monosyllabic vs. trisyllabic pseudowords (Experiment 2). The participants' task was to categorize an /ɑ-aː/ continuum that followed the word lists. The monosyllabic condition was perceived as slower (i.e., fewer /aː/ responses) than the bisyllabic and trisyllabic condition. However, no difference was observed between bisyllabic and trisyllabic contexts. Therefore, while syllable rate is used in perceiving speech rate, other factors, such as fast speech processes, mean F0, and intensity, must also influence rate normalization.

Keywords: Rate normalization, speech rate, syllable rate.

1. INTRODUCTION

Talkers vary in their speech rate, leading to differences in the duration of sentences and the speech sounds in those sentences (e.g., words, syllables, vowels etc.). This poses a problem for speech sounds that are contrasted through temporal information, such as vowels or plosives. Listeners deal with this variability by normalizing vowel perception for the surrounding speech rate [1], [2]. This is a contrastive effect, in which sounds in a fast context are perceived as relatively slow (long), and sounds in a slow context as relatively fast (short). While this effect has been well established, it remains unclear what information listeners use to compute a talker's rate. The present study examined whether the syllable rate (number of syllables per unit time) alone can drive rate normalization.

Previous studies that examined rate normalization have mostly manipulated speech rate by recording a sentence at a normal rate and linearly compressing or expanding the sentence [2], [3], or instructing speakers to produce a natural fast or slow sentence [4], [5]. While these manipulations succeeded in inducing rate normalization, there are many possible cues that could have driven the perception of fast vs. slow tempo, including the number of produced segments [6], [7], the presence of fast speech processes such as reduction [8], but also prosodic cues such as intensity and mean fundamental frequency [9], [10].

Koreman [6] examined whether the number of produced vs. intended syllables/segments affected perceived rate in an explicit rate judgement task. Results showed that sentences with more produced segments (e.g., "He probably said") were perceived as faster than duration-matched sentences with deletions, and thus fewer produced segments per unit time (e.g., "He pro'bly said"). Moreover, sentences with deletions ("He pro'bly said") sounded faster than sentences with the same number of segments but without deletions ("He always said"), and thus a lower number of intended segments. This suggests that listeners track the number of produced and intended segments to compute speech rate. Converging results for syllables, but not for phones, were found by Plug et al. [7]. However, both studies involved explicit rate judgments, not rate normalization.

Reinisch [8] extended these findings to implicit rate normalization. She examined how sentences with or without fast speech processes (deletions and reductions) affected subsequent /a-aː/ perception in German. In contrast to [6] and [7], results showed the exact opposite pattern: duration-matched sentences with fast speech processes (and thus fewer produced segments) resulted in more /aː/-responses than those without, implying that these sentences were perceived as faster. She concluded that the presence of fast speech processes is an indication to listeners of a faster speech rate, in turn influencing subsequent vowel perception.

Despite the diverging results in these previous studies, they all seem to agree that listeners use the number of units (either segments or syllables) per unit time to compute speech rate. However, sentences contain various speech units that differ in their

relative salience. For instance, vowels usually have a greater intensity than consonants, and stressed syllables usually have a greater intensity than unstressed syllables. Considering this difference in salience, do listeners then use all available units alike to compute speech rate, or just the most salient ones?

The present study examined this for syllables, asking: Do listeners use the number of syllables per unit time (i.e., syllable rate) to compute speech rate when strictly controlling for the number of stressed syllables? Alternatively, listeners could use only the stressed syllables (i.e., stress rate) to compute speech rate. We ran two implicit rate normalization experiments using Dutch /ɑ-aː/ targets. In Experiment 1, we compared target perception after duration-matched monosyllabic vs. bisyllabic word lists (i.e., same stress rate, different syllable rate). Experiment 2 compared duration-matched monosyllabic vs. trisyllabic pseudoword lists. If listeners use the syllable rate as the measure for speech rate, monosyllabic word lists should be perceived as two times slower than duration-matched bisyllabic lists and as three times slower than trisyllabic lists. Moreover, if listeners weigh all syllables equally, we might expect a linear relation between the number of syllables in the word lists and the rate normalization effect size.

2. METHODS

2.1. Participants

We recruited 40 native speakers of Dutch from the Radboud University participant pool (Experiment 1: 16 female, 4 male, age range: 18 – 34, M_{age} = 22, SD_{age} = 4.44; Experiment 2: 12 female, 8 male, age range: 21 – 32, M_{age} = 26, SD_{age} = 3.21). All participants gave informed consent and received a monetary reward or course credits for their participation. No participant reported having any speaking or hearing problems.

2.2. Materials

We created word lists in which the first four words served as context to a fifth word (target) that could contain either a short /ɑ/ or a long /aː/. The complete stimulus list has been made available at: https://osf.io/36anr/?view_only=ad1cb3f67e2d4e749620119f7c88c788. All materials were recorded from a female native talker of Dutch.

2.2.1. Context words

For Experiment 1, we selected ten minimally different Dutch word pairs in which one word was monosyllabic, the other bisyllabic with final stress (e.g., *klom,* /klɔm/, 'climbed', vs. *kolom,* /ko.ˈlɔm/, 'column'). Critically, the bisyllabic words differed from the monosyllabic words only in the presence of an unstressed vowel between the first two consonants (1:2 syllable ratio, same stress rate). Next, we set both words of each pair to the same mean duration using PSOLA in Praat [11] (see Figure 1). This was done within word pairs (i.e., *klom* and *kolom* had the same duration, but could differ in duration from another pair).

We then created 12 word lists. Each word list contained either four monosyllabic or four bisyllabic words, with 50 ms of silence between the words, resulting in six monosyllabic lists and six bisyllabic lists. We created six different combinations of the context words, and for every monosyllabic combination of context words, there was a duration-matched bisyllabic version (e.g., *klom, fruit, trein, fluit* vs. *kolom, vooruit, terrein, voluit*).

For Experiment 2, we created ten pairs of pseudowords, following Dutch phonotactics (e.g., *tralk,* /trɑlk/ vs. *tarallok*, /ta.ˈrɑ.lɔk/). In each pair, one pseudoword was monosyllabic, the other trisyllabic with a stressed second syllable (1:3 syllable ratio, same stress rate). Similar to Experiment 1, both pseudowords within a pair were set to the mean duration for the corresponding pair and we again created 12 word lists (6 with monosyllabic items, 6 with trisyllabic items) by concatenating the pseudowords.

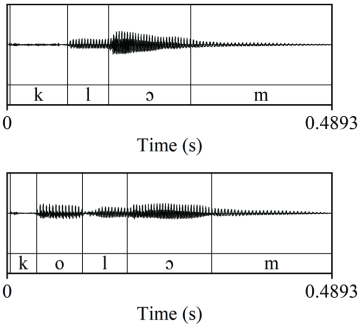

Figure 1: Oscillogram of the context word pair *klom* (top panel) vs. *kolom* (bottom panel)

2.2.2. Target words

We selected six Dutch monosyllabic word pairs that differed only in their vowel, containing either a short /ɑ/ or a long /aː/ (e.g., *stad,* /stɑt/, 'city' or *staat,* /staːt/, 'state').

Next, we required a duration continuum for each word pair, ranging from step 1 (long /aː/) to step 7 (short /ɑ/). Since in Dutch, the /ɑ-aː/ vowel contrast is cued by both temporal and spectral cues, we wanted to control for spectral information by selecting a (close to) ambiguous (i.e., midway between /ɑ-aː/) value for the first and second formant (F1; F2). We

manipulated the F1 and F2 of the vowel based on Burg's LPC method in Praat [11], and set the F1 to 777 Hz and the F2 to 1456 Hz. These values were shown to be perceptually ambiguous in a pretest. For the duration continuum, we manipulated vowel duration using PSOLA in Praat (203 ms (step 1) - 73 ms (step 7); step size = 21.7 ms; the same durations applied to all word pairs).

We then concatenated three middle steps (steps 3, 4, and 5) and the two unambiguous steps (1 and 7) with the word lists to create the stimuli for the experiments (i.e., 'target lists'). Each target word was linked to one word list combination, and all steps of that target word were concatenated with the monosyllabic and bisyllabic versions of that combination (see OSF). The same procedure was also applied in Experiment 2.

2.3. Procedure

The experiment was built and hosted on the Gorilla Experiment Builder (www.gorilla.sc). Participants first performed a headphone screening [12], ensuring that the majority of participants were likely to be wearing headphones.

The main task was a 2AFC experiment in which participants heard the target lists with the target words in list-final position and were instructed to press a button to indicate which word they had heard (e.g., *stad* vs. *staat*).

On each trial, participants were visually presented with a fixation cross in the middle of the screen for 500 ms. Afterwards, we auditorily presented the target lists and at target word offset, two response options (i.e., the two members of the target word pair) appeared on the screen in the middle left and right. Participants were instructed to respond as quickly and accurately as possible from target word offset with button presses ([Z] or [M] for the left or right option, respectively). Response positions were counterbalanced across participants. The next trial started 1 s after their response or after a timed out trial (3 s). Steps 3-5 were repeated 6 times while steps 1 and 7 were repeated twice, serving only as perceptual anchor points. All steps were presented once with a monosyllabic, and once with a bisyllabic word list resulting in 264 trials.

Experiment 2 was almost identical but differed in two ways. First, the 2AFC task was preceded by a familiarization task in which participants were visually presented with the orthography of the pseudowords and auditorily presented with the pseudowords. Participants could repeat the audio up to five times. Second, in the 2AFC task, we visually presented the orthography of the target lists (containing the pseudowords, but not the target word) during the auditory presentation of the target lists. These two changes were added to increase the possibility that the pseudowords were perceived correctly.

3. RESULTS

We excluded trials without a response (Exp1: 0.8%; Exp2: 0.6%). Participants indeed demonstrated ceiling/floor performance on the 'anchor point' steps 1 and 7 (proportion long /aː/ responses: 0.96 and 0.02 (step 1 and 7, Experiment 1), and 0.97 and 0.03 (step 1 and 7, Experiment 2)). Further analysis was restricted to the critical middle steps 3-5. Categorization curves for Experiment 1 and 2 are presented in Figure 2. We ran two separate Generalized Linear Models (GLMM), one for each experiment, with a logistic linking function using the lmerTest package [13] in R [14]. The models tested the binomial categorization responses (long vowel coded as 1; short vowel coded as 0).

The final model for Experiment 1, as obtained through forward modelling, contained the following fixed factors: Rate (categorical predictor with two levels, deviance coded with monosyllabic coded as -0.5 and bisyllabic coded as 0.5), Continuum step (continuous predictor, scaled to z-scores) and their interaction. Next to random intercepts for Participants and Items, the model further included by-Participant random slopes for Rate and by-Item random slopes for Continuum step. Models with a more complex structure failed to converge.

The model revealed a significant effect of Continuum step (β = -1.37, SE = 0.06, z = -22.67, p < .001), confirming that the shorter vowel durations resulted in fewer long /aː/ responses. Second, the model revealed a significant effect of Rate (β = 0.32, SE = 0.11, z = 2.88, p < .01), illustrating that word lists containing bisyllabic context words resulted in a higher proportion of long /aː/ responses compared to word lists with monosyllabic context words. No significant interaction was found.

The final model for Experiment 2 was identical to that for Experiment 1, except that, next to random intercepts for Participants and Items, the model included by-Participant random slopes for Rate and Continuum step.

The model revealed a significant effect of Continuum step (β = -1.59, SE = 0.15, z = -10.64, p < .001), confirming the same effect of vowel duration on /ɑ/-/aː/ responses as in Experiment 1. Second, the model revealed a significant effect of Rate (β = 0.35, SE = 0.11, z = 3.16, p < .01), again illustrating that word lists with trisyllabic context words lead to more long /aː/ responses. However, the size of this effect was numerically not greater than (instead, seemed

comparable to) the monosyllabic vs. bisyllabic contrast in Experiment 1. Indeed, an omnibus model of the combined data showed no interaction between Rate and Experiment ($\beta = 0.01$, $SE = 0.14$, $z = 0.08$, $p = .93$), demonstrating a lack of evidence for a stronger rate effect in Experiment 2 compared to Experiment 1.

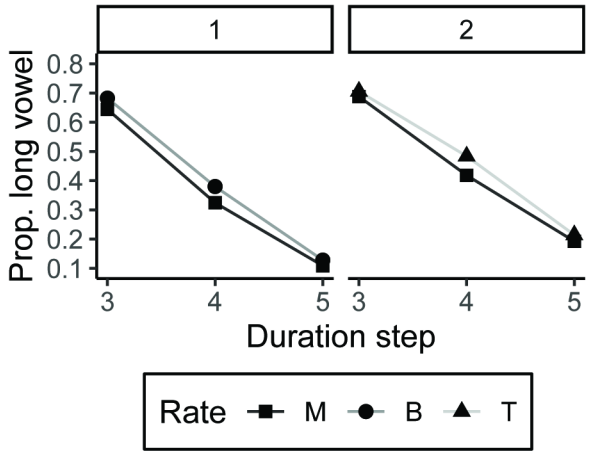

Figure 2: Mean proportion of long /a:/ responses, plotted separately for Experiment 1 (left panel) and Experiment 2 (right panel). Target words were preceded by words lists containing monosyllabic (pseudo)words (M), bisyllabic words (B), or trisyllabic pseudowords (T).

4. DISCUSSION

The present study examined whether the syllable rate is used as a measure of speech rate in an implicit rate normalization task. We found that word lists containing only bisyllabic words led to more /a:/ responses and were thus perceived as faster than duration-matched lists with only monosyllabic words. This illustrates that, despite both words lists having the same stress rate, the number of syllables (i.e., syllable rate) affected vowel perception.

Further, we found that lists containing trisyllabic pseudowords also led to more /a:/ responses compared to lists with monosyllabic pseudowords. However, unexpectedly, the rate effects in Experiment 1 (1:2 syllable ratio) and Experiment 2 (1:3 syllable ratio) were comparable. That is, despite having three times as many syllables in the trisyllabic condition compared to the monosyllabic condition, we did not find a larger rate effect in Experiment 2. This suggests that there is no linear relation between the number of syllables per unit time and the size of the rate effect. These results are in line with Bosker & Ghitza [15], who found that there is an upper limit to the speech rate that can drive rate normalization.

The results of both experiments are consistent with previous explicit rate judgement experiments [6], [7], that found that the number of articulated syllables leads to the perception of a faster tempo. Note that in [6] and [7], the critical comparison was between sentences with or without deletions. Our findings show that, in the absence of a mismatch between the intended and produced number of syllables, listeners also use syllable rate as a measure for speech rate.

An interesting avenue for future research would be to combine the present study with the findings in Koreman [6], Plug et al. [7], and Reinisch [8]. Specifically, if one would compare the bisyllabic lists with new bisyllabic lists in which the vowel of each first syllable is replaced by a reduced vowel (e.g., *kəlom, fəruit, tərein, fəluit*), how would that affect vowel perception? Following [8], the reduced list would be perceived as faster because of the presence of reductions. On the other hand, [6] and [7] would predict the opposite, because the fully articulated bisyllabic words would contain more segments per unit time. Testing such a reduced word list could potentially shed light on the diverging results between [6], [7], and [8].

Moving on to the comparable rate effects in both experiments, the question arises why there is no linear relation between syllable rate and the rate effect. We offer three possible explanations. First, following the conclusions in Bosker & Ghitza [15], syllable rate can only drive speech rate perception to a certain degree. While the present study did not explicitly test where this upper limit lies, it does seem that, with the present stimuli, the 1:2 syllable ratio is close to ceiling. We would predict that if other cues, such as fast speech processes [8] and prosodic cues [7] would be present in the stimuli, we would observe a larger rate effect in Experiment 2. Second, it could be that unstressed syllables carry less weight than stressed syllables in computing speech rate. While unstressed syllables are still important (as is evident from the effect of syllable rate in Experiment 1), listeners might rely more heavily on the stress rate. Third, there is a possibility that the results in the present study could partially be driven by the use of pseudowords in Experiment 2. However, to our knowledge, there is no evidence indicating that non-words would show smaller effects, as rate normalization effects have even been found with non-speech stimuli [1].

In conclusion, the present study provided evidence from an implicit rate normalization task that listeners use syllable rate to compute speech rate. However, syllable rate can only partly account for speech rate normalization, and other cues, such as fast speech processes and prosodic cues, must also play a role in distinguishing different speech rates.

5. ACKNOWLEDGMENTS

This research was funded by the DCC internal round grant, awarded to J.M. and H.R.B. Funding was also received from an ERC Starting Grant (HearingHands, 101040276) from the European Union, awarded to H.R.B. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. We would further like to thank Annelies van Wijngaarden whose voice was recorded for the speech materials used in this study.

6. REFERENCES

[1] H. R. Bosker, "Accounting for rate-dependent category boundary shifts in speech perception," *Atten Percept Psychophys*, vol. 79, no. 1, pp. 333–343, Jan. 2017, doi: 10.3758/s13414-016-1206-4.

[2] L. C. Dilley and M. A. Pitt, "Altering Context Speech Rate Can Cause Words to Appear or Disappear," *Psychol Sci*, vol. 21, no. 11, pp. 1664–1670, Nov. 2010, doi: 10.1177/0956797610384743.

[3] M. Maslowski, A. S. Meyer, and H. R. Bosker, "Eye-tracking the time course of distal and global speech rate effects.," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 46, no. 10, pp. 1148–1163, Oct. 2020, doi: 10.1037/xhp0000838.

[4] 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.

[5] R. S. Newman and J. R. Sawusch, "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, Jan. 2009, doi: 10.1016/j.wocn.2008.09.001.

[6] J. Koreman, "Perceived speech rate: The effects of articulation rate and speaking style in spontaneous speech," *The Journal of the Acoustical Society of America*, vol. 119, no. 1, pp. 582–596, Jan. 2006, doi: 10.1121/1.2133436.

[7] L. Plug, R. Lennon, and R. Smith, "Measured and perceived speech tempo: Comparing canonical and surface articulation rates," *Journal of Phonetics*, vol. 95, pp. 1–15, 2022, doi: 10.1016/j.wocn.2022.101193.

[8] E. Reinisch, "Natural fast speech is perceived as faster than linearly time-compressed speech," *Atten Percept Psychophys*, vol. 78, no. 4, pp. 1203–1217, May 2016, doi: 10.3758/s13414-016-1067-x.

[9] S. Feldstein and R. N. Bond, "Perception of Speech Rate as a Function of Vocal Intensity and Frequency," *Lang Speech*, vol. 24, no. 4, pp. 387–394, Oct. 1981, doi: 10.1177/002383098102400408.

[10] A. C. M. Rietveld and C. Gussenhoven, "Perceived speech rate and intonation," *Journal of Phonetics*, vol. 15, no. 3, pp. 273–285, Jul. 1987, doi: 10.1016/S0095-4470(19)30571-6.

[11] P. Boersma and D. Weenink, "Praat: doing phonetics by computer." Feb. 26, 2019. [Online]. Available: www.praat.org

[12] K. J. P. Woods, M. H. Siegel, J. Traer, and J. H. McDermott, "Headphone screening to facilitate web-based auditory experiments," *Atten Percept Psychophys*, vol. 79, no. 7, pp. 2064–2072, Oct. 2017, doi: 10.3758/s13414-017-1361-2.

[13] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "**lmerTest** Package: Tests in Linear Mixed Effects Models," *J. Stat. Soft.*, vol. 82, no. 13, 2017, doi: 10.18637/jss.v082.i13.

[14] R Core Team, "R: A Language and Environment for Statistical Computing." R Foundation for Statistical Computing, Vienna, Austria, 2020. [Online]. Available: https://www.R-project.org/

[15] H. R. Bosker and O. Ghitza, "Entrained theta oscillations guide perception of subsequent speech: behavioural evidence from rate normalisation," *Language, Cognition and Neuroscience*, vol. 33, no. 8, pp. 955–967, Sep. 2018, doi: 10.1080/23273798.2018.1439179.