Whether long-term tracking of speech rate affects perception depends on who is talking

Merel Maslowski[1,2], *Antje S. Meyer*[1,3], *Hans Rutger Bosker*[1,3]

[1]Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
[2]International Max Planck Research School for Language Sciences, Nijmegen, The Netherlands
[3]Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, The Netherlands

Merel.Maslowski@mpi.nl, Antje.Meyer@mpi.nl, HansRutger.Bosker@mpi.nl

Abstract

Speech rate is known to modulate perception of temporally ambiguous speech sounds. For instance, a vowel may be perceived as short when the immediate speech context is slow, but as long when the context is fast. Yet, effects of long-term tracking of speech rate are largely unexplored. Two experiments tested whether long-term tracking of rate influences perception of the temporal Dutch vowel contrast /ɑ/-/a:/. In Experiment 1, one low-rate group listened to 'neutral' rate speech from talker A and to slow speech from talker B. Another high-rate group was exposed to the same neutral speech from A, but to fast speech from B. Between-group comparison of the 'neutral' trials revealed that the low-rate group reported a higher proportion of /a:/ in A's 'neutral' speech, indicating that A sounded faster when B was slow. Experiment 2 tested whether one's own speech rate also contributes to effects of long-term tracking of rate. Here, talker B's speech was replaced by playback of participants' own fast or slow speech. No evidence was found that one's own voice affected perception of talker A in larger speech contexts. These results carry implications for our understanding of the mechanisms involved in rate-dependent speech perception and of dialogue.

Index Terms: speech rate, rate-dependent speech perception, rate normalization, global context effects, self-produced speech

1. Introduction

Speech rate varies to a surprising degree, both between talkers [1] and within talkers [2]. At the same time, acoustic duration is used to contrast phonemes such as short /ɑ/ and long /a:/ in Dutch. As a consequence of speech rate variation, there is no one-to-one mapping between temporal phonological contrasts and acoustic duration. Therefore, a vowel midway between /ɑ/ and /a:/ may cue either of those vowels, depending on the surrounding speech rate.

In order to map the speech signal onto the intended phonemes, listeners use rate information from the speech context surrounding a temporally ambiguous speech sound. For instance, listeners are more likely to perceive an ambiguous /ɑ, a:/ vowel as /ɑ/ when it is embedded in slow speech, because the slow context makes the vowel sound relatively short [3]. A fast speech context can correspondingly shift perception of the same vowel to /a:/, because fast speech makes it sound relatively long. This influence of context rate on a phonetic boundary between phonemes will be referred to as the rate-dependent phonetic boundary shift (PBS).

Speech rate variation also elicits another rate effect known as the lexical rate effect (LRE) [4] that affects perception of lexical items (e.g., function words) rather than phonemes: Detection of co-articulated function words such as *or* in the sentence *Deena doesn't have any leisure or time* depends on contextual speech rate, with *or* less often being detected in slow speech contexts than in faster speech. The LRE has been argued to be intrinsically different from the PBS [5, 6]. The PBS seems to be insensitive to talker identity, also arising when context and target are produced by different talkers [7]. Moreover, the PBS occurs after non-speech auditory stimuli such as fast and slow tone sequences [6], suggesting that the phenomenon involves general auditory processes. The LRE, on the other hand, requires intelligible speech to be elicited [5].

Most previous work has focused on speech rate effects in *local* contexts; that is, the speech rate of the sentence leading up to a target word [8]. However, the LRE can also be induced by *global* contexts (i.e., long-term tracking of the average speech rate in larger speech contexts) [9]. Baese-Berk et al. [9] compared listener groups that each were exposed to different global speech rates. One group listened to 'neutral' and slowed speech (multiplier 1.2) and another group to slowed speech (multiplier 1.2) and even slower speech (multiplier 1.4). Although groups differed in the *average* rate of the trials to which they were listening, one set of trials was identical across groups (i.e., these trials were produced with the same rate). Baese-Berk et al. observed that global rate biased listeners' perception towards the average rate. That is, the faster the average speech rate across trials, the more function words participants reported hearing on slow trials. Moreover, this was a built-up effect, with function word report being more and more affected by the average speech rate over time.

The present study investigates whether the global effect of speech rate on the LRE generalizes to the PBS, specifically investigating whether a slow or fast global speech rate calculated across multiple talkers influences categorization of the Dutch minimal vowel contrast /ɑ/–/a:/. Two hypotheses were proposed. The Averaging Hypothesis predicted that a high average rate would bias perception of target vowels on slower speech trials *towards* the average, resulting in a relatively *high* proportion of /a:/ responses in slow speech. Similarly, a low average rate would bias perception towards /ɑ/ in faster speech trials. This type of bias would correspond to the global rate effect on the LRE [9]. The Contrasting Hypothesis predicted that a high average rate would induce a bias *away from* the average rate, with a high average rate resulting in an even *lower* proportion of /a:/ responses on slower speech trials. That is, listening to a lot of fast speech would make neutral rate speech sound slow. Such an effect of global rate would be consistent with the contrastive

speech rate effects on the PBS in local contexts.

In conversation, a global speech context often includes one's own speech. As mentioned above, the PBS has been taken to involve general auditory processes. Bosker [10] tested whether self-produced speech in local contexts can also induce the PBS, by comparing /ɑ, a:/ categorization immediately after having produced a sentence at a slow or fast rate oneself. He observed a reduced effect of self-produced speech on phonetic categorization in another talker's speech. However, the effect was restored in another experiment where the recordings from the self-production experiment were played back to the participants. If an effect of global rate on the PBS emerges, a further question involves whether one's own voice is similarly included in computation of the global rate. The current study addresses the question whether one's own speech rate contributes to global rate effects.

This study reports two experiments. Experiment 1 tested whether one talker A's speech rate can affect perception of another talker B in larger speech contexts. Here, a low-rate group listened to talker A speaking at a slow rate and to talker B at speaking at a 'neutral' rate, with the average rate across talkers thus being relatively low (cf. Figure 1; group design). Another high-rate group listened to the same neutral speech from talker B, but to talker A at a fast rate, the average rate thus being relatively high. Perception of talker B was compared between the two listener groups. A group difference in perception of the same speech would indicate an effect of global rate.

Experiment 2 tested whether playback of one's own speech influences perception of another talker's speech in larger speech contexts in the same way as one talker's speech influences perception of another talker (cf. Experiment 1). Therefore, Experiment 2 was identical to Experiment 1, except that talker A's speech was substituted by the participant's own pre-recorded fast or slow speech. In this experiment, a group difference would suggest that listeners calculate a global rate based on all context speech, including their own.

2. Experiment 1

2.1. Method

2.1.1. Participants

A sample of 32 female native Dutch participants ($M_{age} = 22$) with normal hearing was recruited from the Max Planck Institute participant pool and divided into two groups of 16 participants each. All participants gave their informed consent to take part in the study.

2.1.2. Design

Two native speakers of Dutch (one male and one female) were recorded producing a set of eight 24-syllable sentences with one of two Dutch /ɑ/-/a:/ minimal pairs: *takje*/*taakje* (/tɑkjə, ta:kjə/, "twig"/"task") and *stad*/*staat* (/stɑt, sta:t/, "city"/"state"). Each sentence recording was split into a context phrase (cf. Figure 1; light grey background), buffers (Figure 1; white background) and a target word (Figure 1; dark grey background). Using PSOLA in Praat [11], context phrases and buffers (including the consonantal frames of target words) were set to the mean duration of each interval across the two talkers. Context phrases were then rate manipulated through linear expansion (factor of 1.6) and compression (factor of 1/1.6 = 0.625) with PSOLA, whereas buffers and target word consonants had fixed durations.

In Dutch, the vowel contrast /ɑ/–/a:/ is differentiated both temporally and spectrally [12]. Here, we constructed 5-step vowel duration continua ranging from 80 to 120 ms in steps of 10 ms with ambiguous spectral information. The F1 and F2s from both talkers were computed and adjusted with Burg's LPC algorithm as implemented in Praat (male talker: F1 of 764 Hz and F2 of 1261 Hz; female talker: F1 of 728 Hz and F2 of 1327 Hz). Finally, context phrases, buffers, and target words were concatenated, resulting in 240 unique stimulus sentences (8 context phrases × 3 rates × 5 vowel durations × 2 talkers).

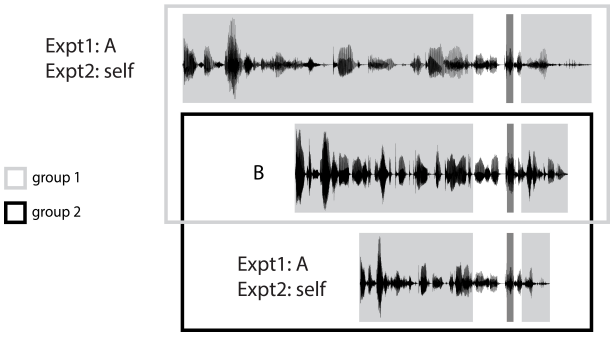

Figure 1: *Experiment design. Each stimulus sentence consists of a context phrase (light grey background), buffers on either side of the target (white background) and the target vowel itself (dark grey background). Context phrases were rate-manipulated (*slow/neutral/fast*), whereas the durations of buffers were fixed. In Experiment 1, group 1 (*low-rate group*) listened to talker A at a slow rate and talker B at a neutral rate (grey box), whereas group 2 (*high-rate group*) listened to neutral rate from talker B, but to talker A at a fast rate (black box). In Experiment 2, talker A was replaced by playback of the participant's own slow or fast speech.*

2.1.3. Procedure

Stimuli were presented in 5 blocks of 80 stimuli, with a random presentation order within each block. Each trial started with a fixation cross on the screen. At auditory stimulus onset, the stimulus sentence also appeared on the screen. After auditory stimulus offset, this screen was replaced by a response screen with two response options (e.g. *stad* and *staat*). If participants did not respond by button press within 4 seconds, a missing response was recorded. One low-rate group listened to slow speech from one talker and neutral speech from the other (talker to rate assignment was counterbalanced). Another high-rate group also listened to neutral speech from one talker, but to fast speech from the other talker.

2.2. Results

Figure 2 presents the categorization data of Experiment 1. Participants reported a higher proportion of /a:/ as vowel duration increased. Each of the four lines in the figure represents a Rate Condition, with line color indicating the between-group condition (*high*/*low* average rate) and line type indicating the within-group condition (*fast*/*neutral*/*slow* trial), resulting in the conditions high|fast, high|neutral, low|neutral, and low|slow. The differences between fast, neutral and slow trials suggest that the proportion of /a:/ responses was higher in fast trials compared to slower trials. The difference between the two solid lines in the center indicates that the high-rate group reported a lower

proportion of /a:/ in neutral speech than the low-rate group, suggesting a contrastive effect of global rate.

Model development and statistical analyses of the categorization data (0.05% missing responses excluded) were performed using a Generalized Linear Mixed Model (GLMM) with a logistic linking function from the `lme4` package [13] in `R` [14]. The predictors in the model included Rate Condition (categorical predictor; intercept is high|neutral), Vowel Duration (continuous predictor; centered and divided by two standard deviations), the interaction between Rate Condition and Vowel Duration, Block (continuous predictor; centered and divided by two standard deviations), the interaction between Rate Condition and Block, and Talker (categorical predictor; sum-to-zero coded) as a control variable. Random intercepts were included for Participant and Item, as well as random slope terms for Vowel Duration and Block, both by Participant and by Item.

Vowel Duration significantly affected the proportion of /a:/ responses ($\beta = 1.145, z = 9.092, p < 0.001$), with the proportion of /a:/ increasing for longer vowel durations. A significant effect of Rate Condition was observed across the three speech rates (high|neutral vs. high|fast: $\beta = 1.846, z = 23.967, p < 0.001$; high|neutral vs. low|slow: $\beta = -1.096, z = -3.409, p < 0.001$); faster speech rates induced higher proportions of /a:/ responses than slower speech rates. Moreover, perception was significantly different in the two groups (high|neutral vs. low|neutral: $\beta = 0.757, z = 2.352, p = 0.019$). The high-rate group who listened to fast and neutral speech reported a lower proportion of /a:/ responses in neutral speech than the low-rate group who listened to neutral and slow speech. The model revealed no main effect of Block ($\beta = -0.180, z = -1.787, p = 0.074$), indicating that performance did not change over the course of the experiment. A significant interaction between high|fast and Block ($\beta = 0.196, z = 2.640, p = 0.008$) was observed, suggesting that the proportion of /a:/ responses in fast speech increased towards the end of the experiment. A significant interaction between high|fast and Vowel Duration ($\beta = -0.467, z = -6.044, p < 0.001$) suggested a ceiling effect in fast speech. Finally, Talker was significant ($\beta = 0.219, z = 4.407, p < 0.001$), indicating that categorization was different for the individual talkers. Altogether, the results indicate that the average rate calculated over multiple talkers affects perception of segments on individual speech trials.

Experiment 2 repeated the experiment with the crucial difference that talker A was now replaced with playback of participants' own voices. If listeners include their own voice in the global speech rate, the results of Experiment 2 should resemble the results of Experiment 1. Alternatively, there should be no difference in perception of neutral rate speech between the high-rate group and the low-rate group.

3. Experiment 2

3.1. Method

3.1.1. Participants

A new sample of 22 native Dutch female participants ($M_{age} = 23$) from the same participant pool as before were tested, and divided into a high-rate group ($N = 10$) and a low-rate group ($N = 12$). More participants were tested, but because participants were recorded and tested on different days, some participants did not return for the test phase. As a consequence, group sizes were mildly unbalanced.

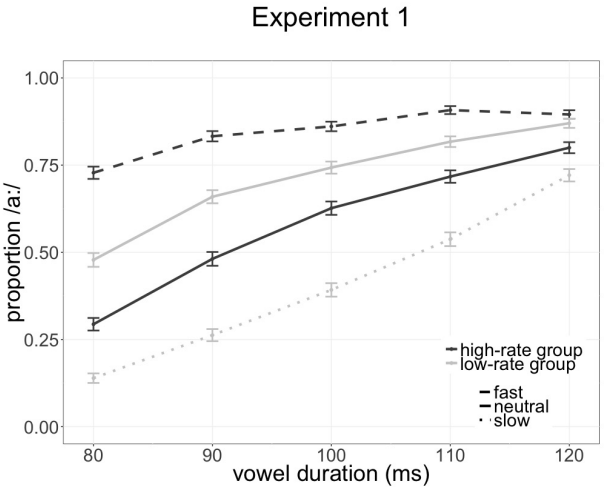

Figure 2: *Average categorization data of Experiment 1. The X-axis indicates Vowel Duration (80 to 120 ms). Rate Condition* fast *is indicated by the dashed line,* neutral *by the solid line, and* slow *by the dotted line. Colors indicate Group, with the high-rate group shown in black and the low-rate group shown in grey. Error bars represent the standard error of the mean.*

3.1.2. Design

The design of Experiment 2 was identical to Experiment 1, except now talker A's speech trials were replaced by 200 sentence recordings from the participants themselves. Participants were instructed to produce the sentences at a pre-specified rate similar to the either fast or slow rate of Experiment 1. In order to avoid influences from participants' own /ɑ/ and /a:/, the target words *takje*/*taakje* and *stad*/*staat* were substituted by *tukje* (/tʏkjə/, "nap") and *stoet* (/stut/, "procession") in these recordings. The self-produced trials were recorded previously in a separate session.

3.1.3. Procedure

The procedure of Experiment 2 was identical to Experiment 1, except that here participants were only prompted to respond after neutral trials from talker B, and not after self-produced speech trials (in which no ambiguous words were present). Instead, the next trial was presented directly after a trial with their own speech.

3.2. Results

Figure 3 presents the categorization data of Experiment 2. Because there was no response after trials in which participants heard their own voices, the lines represent the categorization data of neutral speech only, separately for each group. There seems to be no main effect of group.

A GLMM tested the categorization data of Experiment 2. The fixed effects included Rate Condition (categorical predictor; intercept is high|neutral), Vowel Duration (continuous predictor; centered and divided by two standard deviations), the interaction between Rate Condition and Vowel Duration, Block (continuous predictor; centered and divided by two standard deviations), the interaction between Vowel Duration and Block, and, finally, the control variable Talker (categorical predictor, sum-to-zero coded). Random intercepts were included for Participant and Item, as well as random slope terms for Vowel Du-

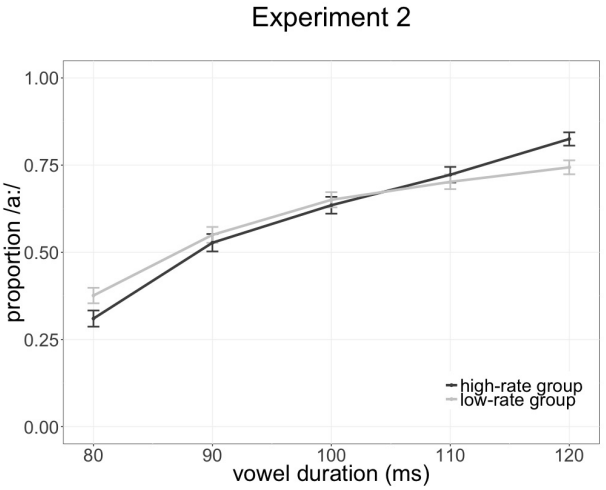

Figure 3: *Average categorization data of Experiment 2. The X-axis indicates Vowel Duration (80 to 120 ms). Colors indicate Group, with the categorization data from the high-rate group shown in black and from the low-rate group shown in grey. Error bars represent the standard error of the mean.*

ration and Block by both random effects.

The proportion of /a:/ responses (0.15% missing responses excluded) was significantly affected by Vowel Duration ($\beta = 1.304, z = 5.081, p < 0.001$), indicating that participants more often reported hearing /a:/ when target vowels had longer durations. Rate Condition did not significantly influence /a:/ categorization ($\beta = -0.230, z = -0.708, p = 0.479$), suggesting that there was no difference in how likely each group was to hear /a:/ in neutral speech. The interaction between Rate Condition and Vowel Duration did not reach significance ($\beta = -0.240, z = -0.846, p = 0.398$), showing that the degree to which /a:/ categorization increased with Vowel Duration was similar between Rate Conditions. No effect of Block was found ($\beta = 0.073, z = 0.881, p = 0.379$), suggesting that performance did not change over time. However, an interaction between Vowel Duration and Block was observed ($\beta = 0.178, z = 3.963, p < 0.001$). The control variable Talker significantly affected categorization ($\beta = 0.890, z = 2.166, p = 0.030$).

4. Discussion

The present study tested whether listeners use the average speech rate calculated across talkers and over a longer period of time in perception of ambiguous speech sounds, resulting in a phonetic boundary shift (PBS). Experiment 1 showed that listeners' perception of temporal phonemic contrasts is modulated by global speech rate, such that one talker's rate influences perception of another talker. However, Experiment 2 revealed that one's own voice is not included in the global rate heard over a longer period of time. Rather, a global speech rate is derived from the speech rate from other talkers only.

The results of Experiment 1 support the Contrasting Hypothesis; the average speech rate of an extended speech context has a contrastive effect on /ɑ/–/a:/ categorization. That is, the presence of *fast* speech biases perception of an ambiguous /ɑ, a:/ in *neutral* rate speech towards short /ɑ/, suggesting that neutral rate speech sounds slower in the context of fast speech. Similarly, neutral rate speech sounds relatively fast in the context of slow speech, as evidenced by a bias towards long /a:/ in the low-rate group.

The effect of global rate on the PBS observed here differs from the global rate effect on the LRE [9] in two important ways. Firstly, the directions of the two effects are inverse; the present global PBS is contrastive in nature, whereas the global LRE is an averaging effect [9]. That is, a slow global rate results in a *higher* proportion of /a:/ responses in neutral speech (i.e., neutral rate sounds fast), whereas it results in a *lower* proportion of function word report (i.e., neutral rate sounds slow).

Secondly, the global PBS arose relatively fast, as Experiment 1 demonstrated no effect of Block. This indicates that categorization was relatively stable over the time course of the experiment. Conversely, the global LRE was a built-up effect on perception, with function word report progressively being affected by the average speech rate. This suggests that the PBS and LRE indeed reflect different underlying mechanisms, as suggested by Bosker [6] and Pitt et al. [5].

As the global rate effect on the PBS is both fast and contrastive in nature, it appears more similar to effects of local context rate on the PBS. Yet, the global rate effect is more than simply an extension of local rate effects to larger contexts. Experiment 2 of the current study demonstrated that global and local effects differ crucially in one important respect. Local speech rate effects have been argued to involve general auditory processes, even arising after one's own speech [10]. However, Experiment 2 demonstrated no effect of self-produced rate in larger speech contexts. Therefore, the results suggest that the global PBS, in contrast to the local PBS, operates at a different perceptual level, differentiating between other talkers' voices and one's own.

5. Conclusions

The findings of the present study are especially interesting in the light of dialogue. Experiment 1 showed that the habitual speech rate of talker A may modulate perception of phonemic contrasts in the speech of talker B. This has consequences for speech comprehension in situations where listeners are exposed to multiple talkers with strongly diverging speech rates. Because perception of a given ambiguous segment is based on a contextual rate that is not relevant for disambiguation of that segment (since it is produced by someone else), such listener situations may lead to misinterpretations. However, Experiment 2 indicates that such misinterpretations are likely to be constrained, as the global rate does not include one's own speech rate. Therefore, disregarding one's own voice in computation of a global rate might prevent communication from failing.

6. References

[1] E. Jacewicz, R. A. Fox, and L. Wei, "Between-speaker and within-speaker variation in speech tempo of american english," *The Journal of the Acoustical Society of America*, vol. 128, no. 2, pp. 839–850, 2010.

[2] J. L. Miller and T. Baer, "Some effects of speaking rate on the production of /b/ and /w/," *The Journal of the Acoustical Society of America*, vol. 73, no. 5, pp. 1751–1755, 1983.

[3] 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, no. 2, pp. 101–116, 2013.

[4] L. C. Dilley and M. A. Pitt, "Altering context speech rate can cause words to appear or disappear," *Psychological Science*, vol. 21, no. 11, pp. 1664–1670, 2010.

[5] M. A. Pitt, C. Szostak, and L. C. Dilley, "Rate dependent speech processing can be speech specific: Evidence from the perceptual disappearance of words under changes in context speech rate," *Attention, Perception, & Psychophysics*, vol. 78, no. 1, pp. 334–345, 2016.

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

[7] 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, 2009.

[8] 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, p. 978, 2011.

[9] 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, no. 8, pp. 1546–1553, 2014.

[10] H. R. Bosker, "How our own speech rate influences our perception of others," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 2017.

[11] P. Boersma and D. Weenink, "Praat: doing phonetics by computer computer program. version 5.4. 09," 2015.

[12] P. Adank, R. Van Hout, and R. Smits, "An acoustic description of the vowels of northern and southern standard dutch," *The Journal of the Acoustical Society of America*, vol. 116, no. 3, pp. 1729–1738, 2004.

[13] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[14] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2014. [Online]. Available: http://www.R-project.org/