How the tracking of habitual rate influences speech perception

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*
**Corresponding author. E-mail address: Merel.Maslowski@mpi.nl*

Abstract

Listeners are known to track statistical regularities in speech. Yet, which temporal cues are encoded is unclear. This study tested effects of talker-specific habitual speech rate and talker-independent average speech rate (heard over a longer period of time) on the perception of the temporal Dutch vowel contrast /ɑ/-/a:/. First, Experiment 1 replicated that slow local (surrounding) speech contexts induce fewer long /a:/ responses than faster contexts. Experiment 2 tested effects of long-term habitual speech rate. One high-rate group listened to ambiguous vowels embedded in 'neutral' speech from talker A, intermixed with speech from fast talker B. Another low-rate group listened to the same 'neutral' speech from talker A, but to talker B being slow. Between-group comparison of the 'neutral' trials showed that the high-rate group demonstrated a lower proportion of /a:/ responses, indicating that talker A's habitual speech rate sounded slower when B was faster. In Experiment 3, both talkers produced speech at both rates, removing the different habitual speech rates of talker A and B, while maintaining the average rate differing between groups. This time no global rate effect was observed. Taken together, the present experiments show that a talker's habitual rate is encoded relative to the habitual rate of another talker, carrying implications for episodic and constraint-based models of speech perception.

Keywords: speech rate, rate-dependent perception, rate normalization, habitual speech rate

Introduction

Humans detect and adapt to statistical regularities in different sensory domains, such as sight, touch, and hearing. In the domain of language, statistical learning has been shown to underlie speech processing and language acquisition (Saffran, Aslin, & Newport, 1996; Saffran, Johnson, Aslin, & Newport, 1999). For instance, the development of phonological categories is sensitive to the probability distributions of acoustic-phonetic cues (Clayards, Tanenhaus, Aslin, & Jacobs, 2008; Maye, Weiss, & Aslin, 2008). The present study examined how listeners track statistical distributions of *temporal* information in speech. It contributes to our understanding of speech perception by showing that listeners adapt to long-term temporal information in a talker-specific manner. We show that a specific talker's habitual speech rate, but not the average speech rate across different talkers heard over a longer period of time, influences subsequent speech perception. These results are important for our understanding of how listeners map variable speech input onto stored phonological representations.

Listeners have been shown to pick up on temporal cues in local speech contexts (e.g., the sentence preceding a target) and use the distributional properties of these temporal cues to adjust subsequent perceptual analysis of speech. This observation will be referred to as rate-dependent speech perception. One manifestation of rate-dependent speech perception is the phonetic boundary shift (PBS). The PBS refers to the fact that contextual speech rate can shift categorization of temporally contrastive phonemes from one phoneme to another (Miller, 1981; Bosker, 2017a; Reinisch, Jesse, & McQueen, 2011; Summerfield, 1981; Wade & Holt, 2005). For instance, perception of the Dutch vowel contrast between short /ɑ/ and long /a:/ is biased towards long /a:/ in a fast, compared to a slower, speech context (Reinisch & Sjerps, 2013). A fast context makes an ambiguous vowel between /ɑ/ and /a:/ sound relatively long (i.e., as /a:/ in *taak* "task"), whereas a slow context makes the same vowel sound short (i.e., as /ɑ/ in *tak* "branch").

The PBS has been shown to be elicited by speech rate variation in the sentence context surrounding the critical segment, even if this local context is produced by another talker than the critical segment (Bosker, 2017b; Newman & Sawusch, 2009). That is, despite the important role of talker variability and talker identity in language processing (Creel & Bregman, 2011; Eisner & McQueen, 2005), the speech rate in a context phrase in one voice can affect phonetic perception of an ambiguous target in another voice. This observation has

been taken to support the idea that the PBS involves general auditory normalization processes that arise early in perception (Bosker, 2017a; Bosker, Reinisch, & Sjerps, 2017; Wade & Holt, 2005).

There is evidence that listeners not only track local temporal information, but also talkers' habitual speech rates (i.e., further removed, more global temporal distributions). For instance, listeners can judge whether certain segmental durations are more or less typical for a given talker (Allen & Miller, 2004; Theodore, Miller, & DeSteno, 2009). Recently, Reinisch (2016) investigated whether knowledge about a talker's habitual speech rate, established by prior exposure, influenced subsequent perceptual processing of that talker's speech. In one experiment, Reinisch first presented participants with a 2-minute dialogue in which one female talker spoke fast and another female talker spoke slowly. After this exposure phase, participants categorized isolated words (i.e., words presented without a speech context) with temporally ambiguous vowels (mid-way between German /a/–/a:/), spoken by the two talkers heard before. Reinisch found that listeners reported more long vowels when evaluating words spoken by the habitually fast talker than by the slow talker, suggesting that listeners adapted their perception of the target vowels based on the habitual rates of the individual talkers in the exposure phase. In a second experiment, participants were presented with the same dialogues as in the first experiment. However, the test phase was different from the first experiment, with the target words from the previous experiment now being embedded in rate-manipulated (local) context sentences. Now only effects of the local context were observed, without any difference between the two talkers. Thus, listeners indeed tracked talkers' habitual rates, adjusting their perceptual phonemic categories accordingly, though the effect of habitual rate was rapidly overridden by effects of more local temporal cues.

The finding that a talker's habitual speech rate influences subsequent perception may be explained by episodic models of speech perception (e.g., Goldinger, 1998). These models hold that each encountered pronunciation of a word is stored, including both linguistic and indexical speech features. Thus, word forms are assumed to be labeled for, for instance, the (slow or fast) speech rate in which it occurred and the talker that produced that particular variant (Pierrehumbert, 2001). Speech perception involves matching incoming acoustic tokens to stored labeled exemplars. Thus, the target words in the categorization task in Reinisch' (2016) first experiment would better match the recently added exemplars from the (fast or slow) talker heard during exposure, explaining the effect of habitual rate observed in Reinisch' Experiment 1.

Another way of conceptualizing the effect of habitual speech rate on perception is within the belief-updating model by Kleinschmidt and Jaeger (2015), where rate-dependent speech perception may be regulated by detection of statistical regularities. This model assumes that listeners have prior beliefs about cue distributions based on previous experience. As listeners process speech, they update their beliefs about the upcoming speech by upweighing or downweighing specific cues. As such, listeners may track statistical distributions of temporal cues that may co-occur with specific situations or with particular talkers, resulting in talker-specific models. These models may then be re-applied to later encounters of that same situation or talker.

Both types of model (episodic and belief-updating) are elegant and powerful frameworks, but neither specifies in detail which cues listeners actually use in specific situations, how they combine and update them, or define the timescale at which temporal cues are tracked/encoded. For simplicity, we adopt the episodic view for further discussion. One debated issue in episodic models is whether more context-specific (signal-extrinsic) indexical properties are encoded and may influence subsequent perceptual processing. Some studies have argued for context-specific, integrated word representations based on evidence that co-occurring non-speech contexts, such as background noise or environmental sounds, affect word learning (Creel, Aslin, & Tanenhaus, 2012), recognition (Pufahl & Samuel, 2014), and memory (Cooper, Brouwer, & Bradlow, 2015). The main goal of the present study was to extend this line of research, investigating which contextual temporal cues are encoded and how sensitive this encoding is to surrounding temporal cues from other talkers.

One specific question that arises from Reinisch (2016) and the frameworks described above, is how talker-specific habitual speech rates are represented by the listener: Is the perceived habitual speech rate of a given talker represented in an absolute manner (e.g., x number of syllables produced by talker A at a given time; i.e., insensitive to the context in which this habitual rate occurred) or is it itself sensitive to surrounding temporal cues produced by others (i.e., influenced by signal-extrinsic temporal cues produced by other talkers)? One might expect that talker A, with an 'average' speech rate, sounds relatively slow if she is heard after a very fast talker. Such a pattern would correspond to contrast effects seen in studies of size or weight estimation, where estimates have been found to depend on the properties of the stimuli judged before (e.g., de Brouwer, Smeets, & Plaisier, 2016). Alternatively, listeners' estimates of speech rate might be tightly linked to specific

talkers and would therefore be rather immune to such cross-talker influences.

First, Experiment 1 was a conceptual replication of previous findings of local rate-dependent PBS (e.g., Reinisch & Sjerps, 2013), testing categorization of the Dutch duration continuum /ɑ/–/a:/. This experiment was conducted to validate the paradigm for investigating rate-dependent speech perception with the constructed stimulus set and to form a baseline for comparison with results of subsequent experiments. Participants listened to two talkers, each producing ambiguous /ɑ/–/a:/ vowels in target words embedded in sentences at three different context rates. We expected that higher contextual speech rates would lead to an increase in the proportion of /a:/ responses, as indeed corroborated by the results.

Experiment 2 was designed to investigate whether or not the perceived habitual speech rate of a talker depends on the speech rate of other talkers heard in the same context. That is, can one talker's habitual speech rate affect the perception of another talker's habitual rate? As in Experiment 1, listeners evaluated an /ɑ/–/a:/ continuum embedded in rate-manipulated context phrases, but now these context phrases were from a male talker and a female talker with distinctly different habitual speech rates. One participant group was exposed to talker A with a 'neutral' habitual speech rate, intermixed with speech from talker B with a fast habitual rate (high average rate; henceforth: high-rate group). Another group listened to the same talker A with a 'neutral' habitual speech rate, but to talker B with a slow habitual speech rate (low average rate; henceforth: low-rate group). Perception of target words embedded in talker A's neutral speech was compared between the high-rate group and the low-rate group.

If different talkers' habitual speech rates are perceived independently of each other, there should not be any difference between the categorization responses of the two groups. That is, talker A's neutral habitual rate would be perceived independent of the temporal cues in talker B's speech, thus exerting the same contextual influence on target word perception across the two groups. However, if the perception of the habitual rate of talker A is sensitive to the habitual rate of talker B, talker A should sound particularly slow in the context of the fast habitual rate of talker B in the high-rate group (and, conversely, particularly fast in the context of the slow habitual rate of talker B in the low-rate group). The result should be a lower proportion of /a:/ responses in talker A's neutral speech in the high-rate group (vs. the low-rate group).

To preview findings, the results of Experiment 2 were consistent with the latter hypothesis: They suggested

that the perceived speech rate of talker A was affected by the speech rate of talker B. It reveals that more contextual (signal-extrinsic) temporal cues are also encoded and influence perceptual processing. This could be explained in one of two ways. Firstly, it could imply that the participants tracked the rates of the two talkers individually, but that the perception of each talker's rate was affected by the other talker's speech rate. An alternative account of the results is that the participants did not track the two talkers individually, but that their perception of the target words depended on the average speech rate across both talkers. Under this account, it is not the fast habitual rate of talker B that made talker A sound slow in the high-rate group, but rather the relatively high average speech rate heard across both talkers.

Discriminating between talker-specific (habitual rate of talker B influenced perception of talker A) and talker-independent (average rate influenced perception of talker A) accounts of the results of Experiment 2 is important for our understanding of whether and which contextual (signal-extrinsic) indexical properties are encoded in speech processing. Therefore, as detailed below, Experiment 3 aimed to distinguish between these accounts, asking whether listeners track temporal cues of speech rates across talkers, or, rather, the temporal cues of distinct talkers.

Experiment 1: Local speech rate effects

Experiment 1 was a validation experiment conducted to replicate the patterns of local rate-dependent PBS typically found in the literature (e.g., Newman & Sawusch, 2009; Reinisch et al., 2011; Reinisch & Sjerps, 2013), in which slowing the preceding context leads to perceiving subsequent ambiguous segments as relatively short and speeding up the context leads to perceiving them as relatively long. Additionally, Experiment 1 aimed to test the magnitude of these local contextual effects in our stimuli, so as to be able to compare them to possibly diverging patterns due to differences in habitual speech rate in the subsequent experiments.

Method

Participants. Native Dutch female participants ($N = 16, M_{age} = 23$) with no hearing, visual or reading deficits were recruited from the Max Planck Institute participant pool. Only female participants samples were obtained, since female participants were easier to recruit and we wanted to keep participants homogeneous

across all experiments. All participants gave their informed consent to participate, as approved by the Ethics Committee of the Social Sciences department of Radboud University (project code: ECSW2014-1003-196). A priori, it was decided to exclude participants with a proportion of /a:/ responses of < 0.1 or > 0.9, as for these participants the stimuli would be insufficiently ambiguous to observe reliable effects of speech rate. None of the participants in Experiment 1 had to be excluded based on this criterion.

Design and materials. A native Dutch male and a female talker were recorded producing multiple tokens of two sets of four sentences (cf. Table 1), each sentence containing 24 syllables. These sentences always contained a member of two /ɑ, a:/ minimal pairs: *takje*/*taakje* (/tɑkjə, ta:kjə/, "twig"/"task") and *stad*/*staat* (/stɑt, sta:t/, "city"/"state"). None of the sentences favored either member of a pair semantically, nor did they contain other instances of the vowels /ɑ, a:/ (e.g., *Toen Evelien gisteren iets onnozels wilde zeggen heeft ze eens stad/staat gezegd tegen Job*, "When Evelien wanted to say something silly yesterday, she said 'city/state' to Job once"). For each sentence, one clear token was selected from each talker. These sentence recordings were then divided into context phrases, buffers, and target words. The target word was one of the aforementioned minimal pairs containing the /ɑ, a:/ contrast (underlined in *Toen Evelien gisteren iets onnozels wilde zeggen heeft ze eens stad/staat gezegd tegen Job*). The three syllables before and one syllable after the target word functioned as buffers (*Toen Evelien gisteren iets onnozels wilde zeggen heeft ze eens stad/staat gezegd tegen Job*)). The speech around the buffers was the context phrase (*Toen Evelien gisteren iets onnozels wilde zeggen heeft ze eens stad/staat gezegd tegen Job*; cf. Table 1).

Context phrases were excised from the recordings on either side of the buffers. First, any long pauses (> 150 ms) in the context phrases were shortened to 150 ms. Subsequently, the durations of the context phrase intervals before and after the target were matched across the two talkers (i.e., set to the mean duration for each interval), using the PSOLA algorithm in Praat (Boersma & Weenink, 2015). Once matched, the context phrases were manipulated in duration through linear expansion (factor of 1.6) and linear compression (factor of $1/1.6 = 0.625$) with PSOLA, resulting in three rate conditions: *fast*, *neutral* (no further rate manipulation), and *slow*.

The buffers around the target words served to control for effects of adjacent duration information. Buffers were extracted from the original recordings and were matched (set to the mean) in duration for the two

talkers. After this, no time compression or expansion was performed, such that the duration of buffers was fixed irrespective of the rate condition of the context phrase.

To create the target words, /ɑ, a:/ vowel continua were made. In Dutch, the /ɑ, a:/ vowel contrast is acoustically differentiated by both temporal and spectral information (Adank, Van Hout, & Smits, 2004). Therefore, duration continua with spectrally ambiguous F1s and F2s were created. First, one clear long vowel /a:/ was extracted for each talker. Based on the mean durations of /ɑ/ ($M_{male} = 61$ ms; $M_{female} = 56$ ms) and /a:/ ($M_{male} = 147$ ms; $M_{female} = 123$ ms) in our recordings, duration continua ranging from 80 to 120 ms in five steps of 10 ms were made with PSOLA. Subsequently, spectral manipulations were performed based on Burg's LPC method (implemented in Praat), with the source and filter models estimated automatically from the selected vowel. The filter coefficients of the vowels were then adjusted, and thereafter recombined with the source model, resulting in spectral continua of F2. The F1s in the continua were set at constant values, fixed at each talker's mean in their own production (male: 764 Hz; female: 728 Hz). Because /ɑ/ and /a:/ spectrally mainly differ in F2, the F2 values were based on an online pretest (2AFC), in which twelve participants had to classify a set of vowels for each of the two talkers (5 F2 values × 5 vowel durations × 2 talkers = 50 unique stimuli). For each talker, one maximally ambiguous F2 was selected (male: 1261 Hz; female: 1327 Hz) and applied to the duration continuum. For the resulting temporally and spectrally manipulated vowels, the intensity and pitch contours were controlled. The consonantal frame for the vowels was fixed, such that only the vowel of the target word was manipulated.

Finally, context phrases, buffers, and target regions were concatenated, resulting in a stimulus set of 240 unique stimuli (8 context phrases × 3 rates × 5 vowel durations × 2 talkers).

Procedure. Stimulus presentation was controlled by Presentation software (v16.5; Neurobehavioral Systems, Albany, CA, USA). The experiment started with a practice round, in which each of the eight different sentences occurred once in one of the three speech rate conditions. Until the offset of each auditory stimulus, a fixation point was shown on the screen. Then, this screen was replaced by another screen with two response options (e.g., *takje* and *taakje*), after which participants had 4 seconds to indicate which word they had heard. For the word shown on the left of the screen they pressed "1" and for the word shown on the right side of the screen they pressed "0". The position of the response options on the screen was counterbalanced across

participants. If no response was given within 4 seconds, a missing response was recorded. The 240 stimuli were presented to each participant once in a randomized order. One session lasted approximately 25 minutes.

Results and Discussion

Figure 1 summarizes the categorization data (proportion /a:/ responses) of Experiment 1. The figure shows that participants reported a higher proportion of /a:/ when the target vowels had longer durations. The difference between the three lines shows that the proportion of /a:/ responses increased with contextual local speech rate, such that target vowels embedded in fast context phrases received a higher proportion of /a:/, compared to target vowels embedded in slower context phrases.

The categorization data (0.1% missing responses excluded) were tested using a Generalized Linear Mixed Model (GLMM) with a logistic linking function from the `lme4` package (Bates, Mächler, Bolker, & Walker, 2015) in `R` (R Core Team, 2014). The predictors included in the model were Context Rate (categorical predictor; intercept is neutral), Vowel Duration (continuous predictor; centered and divided by one standard deviation), and their interaction. Additionally, Talker (categorical predictor; sum-to-zero coded) was added as a fixed effect to control for differences between the male and the female talker. Random intercepts for Participant and Item were included, as well as random slopes for Context Rate and Vowel Duration, both by Participant and by Item. Slope terms for the interaction between Context Rate and Vowel Duration were dropped, because the corresponding model failed to converge.

The proportion of /a:/ responses significantly increased with vowel duration ($\beta = 0.832, z = 5.180, p < 0.001$). Moreover, the proportion of /a:/ responses significantly increased for fast context phrases ($\beta = 1.027, z = 5.577, p < 0.001$), and significantly decreased for slow context phrases ($\beta = -1.010, z = -4.551, p < 0.001$) relative to the neutral condition that was mapped onto the intercept. This indicates that the faster the context speech rate, the higher the probability of hearing /a:/. A significant effect of Talker was also observed ($\beta = 0.317, z = 3.713, p < 0.001$), with a higher proportion of /a:/ responses for the female talker. The interaction between Context Rate and Vowel Duration did not reach significance (neutral vs. fast $\beta = 0.029, z = 0.236, p = 0.814$; neutral vs. slow $\beta = 0.070, z = 0.639, p = 0.523$).

These results demonstrate that /ɑ, a:/ categorization was influenced by the local rate-manipulated context

phrases, with fast context phrases inducing a perceptual bias towards long /a:/ and slow phrases inducing a perceptual bias towards short /ɑ/. The results replicate speech rate effects reported in previous literature (cf. Bosker, 2017a; Reinisch & Sjerps, 2013), supporting the validity of the paradigm and stimuli to investigate rate-dependent speech perception. The results of this experiment served as a baseline for the evaluation of results in subsequent experiments.

Experiment 2: Inter-talker variation

Experiment 2 aimed to evaluate whether talkers' long-term habitual speech rates are perceived in an absolute manner or relative to other talkers. This was done by comparing listeners' categorization responses to vowels mid-way between /ɑ/ and /a:/ embedded in speech from two talkers with distinct habitual speech rates. The high-rate group of participants listened to talker A producing speech at a neutral rate and to talker B producing speech at a fast rate, whereas the low-rate group listened to the same neutral rate speech from talker A, but to talker B speaking slowly. If the perception of the 'neutral' habitual speech rate of talker A is influenced by the habitual rate of talker B, we would expect differential perception of talker A's speech in the two groups.

Method

Participants. Native Dutch female participants ($N = 38, M_{age} = 22$) who had not participated in Experiment 1 were recruited according to the same selection criteria and from the same participant pool as in Experiment 1. Participants gave their informed consent to participate. Data from 6 participants were excluded, because their responses were outside the set performance range described in Experiment 1, resulting in two pseudo-random groups of each 16 participants.

Design and materials. The same materials were used as in Experiment 1.

Procedure. The procedure was similar to that of Experiment 1, except that now two groups of participants were exposed to different parts of the stimulus set. The high-rate group listened to neutral speech from talker A intermixed with fast speech from talker B (i.e., the average speech rate was high). The low-rate group

listened to the same neutral speech from talker A, but to slow speech from talker B (the average speech rate being low) (see Figure 2). Rate assignment to talker was counterbalanced across participants, such that the female talker was talker A half of the time. Therefore, each participant listened to 80 of the 240 unique stimuli (8 context phrases × 5 vowel durations × 2 rates/talkers). Five blocks of these 80 stimuli were presented to each participant (presentation order within block randomized). As in Experiment 1, each trial started with a fixation point on the screen. At stimulus onset the stimulus sentence appeared on the screen, with a question mark between square brackets in place of the target word (e.g., *Peter fluisterde Ilse iets verkeerd in en toen hoorde Ilse het [?] gezegd worden.*). At stimulus offset, this screen was replaced by the same response screen as in Experiment 1, where participants had 4 seconds to indicate which word they had heard at the position of the question mark. One session lasted for a duration of approximately 40 minutes in the high-rate group and 50 minutes in the low-rate group.

Results and Discussion

Figure 3 represents the categorization data of Experiment 2. Participants reported a higher proportion of /a:/ for vowels with longer durations. The difference between the three line types indicates that participants responded differently to the same vowel depending on the local context speech rate. The difference between the two solid lines in the middle suggests that the perception of vowels embedded in neutral speech was influenced by long-term temporal cues.

A GLMM tested the binomial responses of Experiment 2 (0.05% missing responses excluded). A new variable Rate Condition was created, merging the between-group condition (high/low average rate) with the within-group condition (fast/neutral/slow trial). Rate Condition consisted of four contiguous levels of rate, corresponding to the four lines represented in Figure 3, namely high_fast, high_neutral, low_neutral, and low_slow (where the between-group factor is shown on the left of the underscores and the within-group factor is shown on the right of the underscores). The fixed effects included were Rate Condition (categorical predictor; intercept is high_neutral), Vowel Duration (continuous predictor; centered and divided by one standard deviation), the interaction between Rate Condition and Vowel Duration, Block (continuous predictor; centered and divided by one standard deviation), the interaction between Rate Condition and Block, and Talker

(categorical predictor; sum-to-zero coded) as a control variable. The random effect structure consisted of intercepts for Participant and Item and random slope terms for Vowel Duration and Block by both random effects. Because each participant only responded to two out of four levels in Rate Condition, no random slope terms for this predictor were included.

The proportion of /a:/ responses significantly increased with vowel duration ($\beta = 1.145, z = 9.092, p < 0.001$), with longer vowels more often being heard as the long vowel /a:/. Furthermore, perception differed significantly across the three context speech rates (high_fast vs. high_neutral: $\beta = 1.846, z = 23.967, p < 0.001$; low_slow vs. high_neutral $\beta = -1.096, z = -3.409, p < 0.001$). The target vowels heard in fast context phrases were perceived as longer than those in neutral context phrases, and vowels in neutral contexts were heard as longer than vowels embedded in slow speech. More importantly, performance in low_neutral vs. high_neutral contexts was also significantly different (i.e., a between-group effect; $\beta = 0.757, z = 2.352, p = 0.019$), with vowels embedded in talker A's neutral speech more often being perceived as /a:/ when participants also listened to slow speech from talker B (compared to fast speech from talker B).

Order effects were analyzed by Block, as the randomized trial structure did not permit more fine-grained analyses. There was no significant main effect of Block ($\beta = -0.180, z = -1.787, p = 0.074$), providing no evidence that overall performance changed over time. Moreover, the difference in performance between Rate Conditions low_neutral and high_neutral across the two groups was already visually present in Block 1. However, the interaction between Block and the contrast between Rate Conditions high_fast and high_neutral was significant ($\beta = 0.196, z = 2.640, p = 0.008$), indicating that the difference between high_fast and high_neutral became slightly larger in the high-rate group in later blocks. The interaction between Vowel Duration and the contrast between Rate Conditions high_fast and high_neutral was significant ($\beta = -0.467, z = -6.044, p < 0.001$), possibly due to a ceiling effect in fast speech. The model also accounted for differences between talkers, with a significantly higher proportion of /a:/ responses for the female talker ($\beta = 0.219, z = 4.407, p < 0.001$).

Also, visual comparison of Figure 1 and Figure 3 seems to indicate that fast speech was perceived as faster in Experiment 2 than in Experiment 1 (i.e., higher proportion of /a:/ responses for the fast condition in Experiment 2 compared to Experiment 1). Similarly, slow speech seems to receive a lower proportion of

/a:/ in Experiment 2 than in Experiment 1, the values in Experiment 2 consequently being more extreme. We compared Experiment 1 and 2 by subsetting the responses to target vowels embedded in fast and slow speech only. A GLMM comprising Context Rate (sum-to-zero coded: slow coded as -0.5, fast as 0.5), Experiment (sum-to-zero coded: Experiment 1 coded as -0.5, Experiment 2 as 0.5), Vowel Duration, and Talker, as well as the interaction between Context Rate and Experiment revealed a main effect of Context Rate ($\beta = 2.481, z = 11.023, p < 0.001$). This showed, once more, that there was a difference in vowel categorization between Context Rates fast and slow across the two experiments. The main effect of Experiment was not significant ($\beta = 0.386, z = 1.029, p = 0.303$), suggesting that, averaging across Context Rates, the proportions of /a:/ responses in Experiment 1 and in Experiment 2 were comparable. However, the interaction between Experiment and Context Rate was significant ($\beta = 1.135, z = 2.535, p = 0.011$), indicating that the difference in /a:/ categorization between fast and slow speech was more extreme in Experiment 2, compared to that difference in Experiment 1. Target vowels were less often heard as /a:/ in fast speech in Experiment 1 than in Experiment 2, and they were more often heard as /a:/ embedded in slow speech in Experiment 1 than in Experiment 2.

In sum, the results of Experiment 2 show that talker A's neutral speech received a lower proportion /a:/ responses in the high-rate group than in the low-rate group, indicating that A's speech sounded slow when B was faster, but fast when B was slower. Likewise, comparison of Experiment 1 and Experiment 2 showed that perception of B's speech was affected by the speech rate of A, with B's fast (or slow) speech sounding even faster (or slower) in Experiment 2.

These results suggest that listeners track habitual speech rate not in an absolute, but in a relative manner: The perception of talker A's habitual speech rate is influenced by surrounding talkers' habitual rates. Alternatively, one may argue that the perception of talker A's speech was affected by the average (high/low) speech rate across talkers, rather than the habitual speech rate of talker B. Which of these two accounts best represents how listeners encode long-term rate was investigated in Experiment 3.

Experiment 3: Intra-talker variation

Experiment 2 found a discrepancy between groups in the perception of talker A. This could either be due to listeners tracking habitual rates talker-specifically (fast talker B affects perception of the speech rate of talker A) or due to listeners tracking the average rate across talkers (high average speech rate across talkers affects perception of the speech rate of talker A). To decide between these accounts, Experiment 3 tested whether the speech rate effect found in Experiment 2 would persist when talkers' speech rate distributions were comparable (as opposed to Experiment 2, where talkers had distinct habitual speech rates). Therefore, in Experiment 3 (similar to Experiment 2), a high-rate group listened to fast and neutral speech and a low-rate group to neutral and slow speech. Whilst Experiment 2 manipulated inter-talker rate variation (e.g., talker A was neutral and talker B was fast), Experiment 3 used intra-talker rate variation (e.g., talker A and talker B were both neutral and fast). The average speech rate was still high (low) in the high-rate (low-rate) group, as in Experiment 2. However, the distinction between the habitual speech rates of the two talkers was removed. If listeners track rates talker-independently (i.e., average rate across talkers), the results of Experiment 3 should mirror those from Experiment 2. Alternatively, if listeners track temporal cues talker-specifically (i.e., specific talkers' habitual rates), no difference between the two groups in the perception of neutral trials would be predicted in Experiment 3.

Method

Participants. Native Dutch female participants ($N = 40, M_{age} = 21$) who had not participated in the previous experiments were recruited from the same participant pool as before and gave their consent to participation. Data from eight participants were excluded on the basis of the criteria described in Experiment 1. The remaining participants formed two pseudo-random groups of 16 participants each.

Design and materials. The same materials were used as in the previous experiments.

Procedure. The procedure was identical to that of Experiment 2, except that participants now listened to both talkers speaking at two different rates (i.e., intra-talker variation instead of inter-talker variation). A high-rate group listened to neutral speech from both talker A and talker B intermixed with fast speech from

both talkers. Similarly, a low-rate group listened to neutral and slow speech from both talkers. As a result, each participant listened to 160 unique stimuli (8 context phrases $\times$ 5 vowel durations $\times$ 2 rates $\times$ 2 talkers). These stimuli were presented in a randomized order in each of three blocks. One session lasted for a duration of approximately 50 minutes in the high-rate group and 60 minutes in the low-rate group.

Results and Discussion

Figure 4 presents the categorization data of Experiment 3. Participants reported a higher proportion of /a:/ with increasing vowel duration. The difference between the three line types indicates that there is an effect of local (sentence) speech rate. However, there is no difference between the two solid lines in the middle of the graph representing neutral speech, suggesting that there is no effect of the average (high or low) long-term rate.

A GLMM tested the categorization data of Experiment 3 (0.9% missing responses excluded) to analyze whether the average speech rate affects perception when intra-talker rate variation is present. The model included the predictors Rate Condition (categorical; intercept is high_neutral), Vowel Duration (continuous; centered and divided by one standard deviation), Block (continuous; centered and divided by one standard deviation), and Talker (categorical; sum-to-zero coded). No interactions between predictors were included in the final model, as more complex models including the interactions did not explain the data significantly better. Random intercepts were included for Participant and Item with slopes for all predictors except Talker (control variable) and Rate Condition (as each participant was only exposed to half of the levels of this predictor).

The GLMM revealed a significant effect of Vowel Duration ($\beta = 1.012, z = 8.964, p < 0.001$), with longer vowels more often being perceived as /a:/. The proportion of /a:/ responses was also significantly affected by context speech rate (high_fast vs. high_neutral: $\beta = 0.954, z = 15.302, p < 0.001$; low_slow vs. high_neutral: $\beta = -1.125, z = -4.277, p < 0.001$). However, there was no significant difference between the two groups in perception of vowels embedded in neutral rate (low_neutral vs. high_neutral: $\beta = -0.139, z = -0.529, p = 0.597$). Block did not significantly affect the proportion of /a:/ responses ($\beta = 0.045, z = 0.744, p = 0.457$), indicating that performance did not change over the course of the experiment. Finally, Talker had a significant

effect on performance, with vowels from the female talker more often being reported as /a:/ than vowels from the male talker ($\beta = 0.115, z = 2.742, p = 0.006$).

To further verify that (the absence of) the group effect in this experiment was different from the effect in Experiment 2, we ran another analysis on a subset containing only the neutral rate data from both experiments. The GLMM contained the fixed effects Rate Condition (sum-to-zero coded: low_neutral as -0.5, high_neutral as 0.5), Experiment (sum-to-zero coded: Experiment 2 coded as 0.5, Experiment 3 as -0.5), Vowel Duration, Talker, and the interaction between Rate Condition and Experiment (note that Block was excluded, because block length differed across the two experiments). The random effects included Participant and Item. The main effect of Rate Condition was not significant ($\beta = -0.408, z = -1.720, p = 0.085$), suggesting that there was no consistent difference across both experiments between the high-rate groups and the low-rate groups in perception of neutral speech. There was also no main effect of Experiment ($\beta = 0.193, z = 0.810, p = 0.416$), suggesting that, averaging across Rate Conditions, there was no difference between Experiment 2 and Experiment 3 in /a:/ categorization. However, the model showed a significant interaction between Experiment and Rate Condition ($\beta = -0.959, z = -2.02, p = 0.043$), indicating that no group difference in the perception of neutral speech was present in Experiment 3, whereas it was present in Experiment 2. These analyses demonstrate that there was no *overall* effect of Experiment, yet specifically the group effect (i.e., comparison of low_neutral and high_neutral) was present in Experiment 2, but absent in Experiment 3.

In sum, the results of Experiment 3 showed that the group effect in Experiment 2 disappeared when the two talkers' speech rates had similar distributions. This difference between Experiments 2 and 3 suggests that listeners track long-term rate distributions in a talker-specific manner (i.e., talkers' habitual rates), as opposed to tracking rates in a talker-independent manner (i.e., average speech rate across talkers). The results of this experiment therefore suggest that talkers' habitual rates were the driving factor for the group effect observed in Experiment 2.

General Discussion

Three experiments were performed to test how listeners track long-term temporal cues in speech from different talkers. Experiment 1 aimed to replicate the earlier finding that variation in speech rate in the *local* context (i.e., the surrounding sentence context) induces a phonetic boundary shift (PBS) (e.g., Reinisch & Sjerps, 2013). Results indicated that listeners were more likely to categorize an ambiguous vowel mid-way between /ɑ/ and /a:/ as a long vowel /a:/ when it was embedded in fast context phrases, but as a short vowel /ɑ/ when embedded in slower context phrases.

In Experiment 2, we investigated whether or not perception of a talker's habitual speech rate was influenced by the habitual speech rate of another talker. In this experiment, a high-rate group listened to ambiguous target vowels (mid-way between /ɑ/ and /a:/) produced by talker A speaking at a neutral rate and talker B speaking at a fast rate, whereas a low-rate group listened to ambiguous target vowels produced by neutral talker A and *slow* talker B. That is, the two groups listened to the same neutral rate sentences (i.e., local rate cues) from talker A, yet they differed in the habitual speech rate of talker B. The results indicated that A's neutral speech rate sounded fast (as evidenced by a higher proportion of /a:/ responses) in the context of a slow talker B. This suggests that a listener's perception of a talker's habitual speech rate is sensitive to the habitual speech rate of another talker heard in the same context.

Because the two groups in Experiment 2 differed in both the speech rate of talker B (fast/slow) and the average speech rate across the two talkers (high/low), the difference in perception of talker A between the two groups could either be due to listeners tracking individual talkers' habitual speech rates (i.e., talker-specificity), or to listeners tracking the average speech rate across talkers (i.e., talker-general). This latter account would be in line with studies demonstrating effects of the preceding average stimulus rate on perceived durations, for instance in the field of auditory perception (perceived tempo judgments; Jones & McAuley, 2005; McAuley & Miller, 2007). Experiment 3 was conducted to differentiate between these two possibilities. The crucial difference to Experiment 2 was that participants now heard both talkers speaking at two rates, thus removing the difference in habitual speech rates of talker A and B, with only the average rate differing between groups. Now, the group effect of Experiment 2 disappeared.

The findings of the present study contribute to our understanding of how listeners adapt to talkers'

habitual rates. It complements Reinisch (2016), who investigated whether listeners tracked talkers' habitual rates in a conversation. After listening to a two-minute dialogue between two female talkers with distinct habitual rates in an exposure phase, participants in the test phase categorized vowels in ambiguous isolated words (i.e., without local sentence contexts) from either talker. Reinisch observed an effect of habitual rate on the perception of these isolated words when no other (local) rate information was available. Considering these findings in light of the results of our Experiment 2, the habitual rate effect in Reinisch' experiment may actually have been enhanced by the presence of another talker with a distinctly different habitual rate (i.e., the fast talker sounded particularly fast in the context of the co-occurring slow talker).

Furthermore, in Reinisch' (2016) second experiment, the test phase involved categorization of ambiguous words embedded in rate-manipulated context sentences. In that experiment, talker-specific habitual rate information no longer had an effect on perception. This observation may be interpreted in relation to the fact that we found no long-term rate effect in our Experiment 3, where there was considerable within-talker rate variation. That is, the absence of an effect of habitual rate in Reinisch' second experiment may be explained by the greater within-talker rate variability induced by the rate-manipulated sentences in the test phase (relative to her first experiment).

Another study relevant to the question of how long-term rate distributions affect speech perception and particularly pertinent to our Experiment 3, was conducted by Baese-Berk et al. (2014). This study investigated a rate-dependent effect on speech perception known as the Lexical Rate Effect (LRE). The LRE concerns function word perception: Heavily coarticulated function words like *or* in the phrase *Deena doesn't have any leisure or time* are less often detected when the surrounding stretches of speech are perceived as slow (Dilley & Pitt, 2010). Similarly, function words never originally spoken can be perceived in fast speech. In contrast to the absence of an effect of average rate in our Experiment 3, Baese-Berk et al. (2014) found that the LRE was sensitive to the average rate heard over a longer period of time: The faster the average rate of speech presented over the course of an hour, the more function words participants reported in context phrases that were slower than this average speech rate; that is, slower rates now sounded *less* slow.

There are several differences between our Experiment 3 and the study by Baese-Berk et al. (2014) that could be responsible for the different outcomes. One potentially important difference concerns the different

rates that were compared in each study. In the present experiments, differences between rates were large and salient (ratios 0.625 for fast, 1 for neutral, and 1.6 for slow), whereas successive rates in Baese-Berk et al.'s study differed by only 20%. Maybe listeners are more likely to average speech rates that are more similar to one another than speech rates that are very far apart. For instance, Jones and McAuley (2005) investigated how time judgments of tones are affected by long-term contexts with the same mean rate but different rate distributions (wide vs. narrow), and found lower accuracy scores for wider-range distributions. Additionally, they observed that more errors were made when the local rate change between two trials was large than when it was smaller. This suggests that averaging may be more likely over relatively small differences.

Another difference is that the current study focused on segmental ambiguities in content words, whilst Baese-Berk et al. (2014) investigated a lexical effect, the perception of function words. Pitt, Szostak, and Dilley (2016) have argued that the PBS and the LRE are qualitatively different from each other. Consistent with this view, the PBS has been found to be triggered by non-speech auditory stimuli (such as pure tones; Bosker, 2017a), whereas the LRE is elicited by intelligible speech contexts only (Pitt et al., 2016). Bosker (2017a) has speculated that the difference between the two phenomena may lie in the levels of processing on which they operate, with the PBS being a sublexical and domain-general process and the LRE being a lexical domain-specific process. Therefore, the conflicting results found in the present study and Baese-Berk et al. could also be related to the perceptual locus of the two effects.

The present study, together with Reinisch (2016), demonstrates that talkers' habitual rates can influence speech perception, but only when the rate variation within a particular talker is relatively small. This may be due to listeners having limited capacity to track rate variability within talkers. It is as yet unclear what amount of within-talker variability is allowed before the tracking of talkers' habitual rates breaks down. Considering that rate variation tends to be larger within than between speakers (Miller, Grosjean, & Lomanto, 1984; Quené, 2008), the contribution of tracking of habitual rate to comprehension in natural conversation may have limited impact. Nevertheless, these findings do carry implications for different models of speech perception, including episodic and constraint-based models.

Episodic models of speech perception assume detailed representations (exemplars) based on linguistic experience including rich acoustic detail (Bybee, 2006), possibly in addition to more abstract representations

(e.g., McQueen, Cutler, & Norris, 2006). Detailed exemplars also encode talker-specific information about, for instance, habitual speech rate (Goldinger, 1992; Pisoni, 1993), which may be used in encounters of the known talkers (Johnson, 1997; Pierrehumbert, 2001). The encoding of talker characteristics could explain the differences in perception between the male and female talkers in our experiments; tokens from the two talkers may be labeled differently due to previous experience with other males and females.

Considering the present findings in light of episodic models, our results suggest that these models should include labels for more contextual (signal-extrinsic) temporal cues. As such, this study contributes to the debate about whether (and which) context-specific signal-extrinsic indexical properties of spoken words are encoded during perceptual processing. Not only can contextual factors such as background noise and environmental sounds influence speech perception (Cooper et al., 2015; Creel et al., 2012; Pufahl & Samuel, 2014), but the larger conversational context (i.e., the rate of other surrounding talkers) may also be stored. In turn, this would allow for the possibility that the perception of the habitual rate of one talker is influenced by the perception of the habitual rate of another talker.

The results can also be interpreted within Kleinschmidt and Jaeger's (2015) belief-updating model of perceptual adaptation. The patterns of results seen in our experiments could be due to the beliefs that listeners had about the cue distributions in the speech signal for each talker. Prior to the experiment, listeners had a talker-general model of Dutch based on previous experience and expectations built upon this experience. When they participated in our experiments, their perception of the two unfamiliar talkers was updated, integrating the new experiences from the experiment. As listeners were processing incoming speech from a particular talker, they updated their beliefs about the upcoming speech from that talker. When the listener observed that talkers spoke at stable habitual rates (Experiment 2), they upweighted talker-specific cues, relying on a specific model for each talker. However, the beliefs about these talker-specific cues were partly based on the speech from another talker (e.g., the belief that one talker must be fast, as the other talker is slower). In Experiment 3, the listener observed that talkers' rate distributions were comparable. Therefore, the listener either grouped the two talkers together, downweighting talker-specific cues (with the listener henceforth relying on the same general model for both talkers), or the listener relied on a specific model for each talker, with the two talker models being very similar (with regard to speech rate). The latter

option may account for the consistent differences found in perception of our male and female talker (i.e., higher proportions of /a:/ responses for the female talker than for the male talker).

The current study shows effects of temporal cues in the local surrounding context and effects of temporal cues in (more long-term) global contexts. Whereas effects of local contexts operate independent from talker-identity (i.e., when a sentence in one voice influences perception of a target word in a different voice; Bosker, 2017b; Newman & Sawusch, 2009), global rate effects seem to be sensitive to the habitual rates of particular talkers (cf. our Experiment 2; Reinisch, 2016). This suggests that these two types of context effects dissociate, indicative of a hierarchical cognitive framework with at least two stages. This would be in line with a recent proposal by Bosker et al. (2017), who have proposed a two-stage model of (temporal and spectral) normalization processes in speech perception. The first stage involves automatic general auditory mechanisms, operating early in perception, unaffected by attentional modulation (e.g., talker segregation; cognitive load; speech vs. non-speech). A second stage involves cognitive (rather than perceptual) adjustments on the basis of higher-level influences, such as comparing a target sound to its expected realization given a certain context (e.g., a particular talker). We speculate that the effects of local surrounding context operate at the first (automatic, general-auditory) stage, whereas global rate effects would operate at the second stage, involving later cognitive adjustments. Future experiments may further test this framework by examining the time course of local and global rate effects.

References

Adank, P., Van Hout, R., & Smits, R. (2004). An acoustic description of the vowels of Northern and Southern Standard Dutch. *The Journal of the Acoustical Society of America*, *116*(3), 1729–1738.

Allen, J. S., & Miller, J. L. (2004). Listener sensitivity to individual talker differences in voice-onset-time. *The Journal of the Acoustical Society of America*, *115*(6), 3171–3183.

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

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

Boersma, P., & Weenink, D. (2015). *Praat: doing phonetics by computer computer program. Version 5.4. 09.* Retrieved from `http://www.praat.org/`

Bosker, H. R. (2017a). Accounting for rate-dependent category boundary shifts in speech perception. *Attention, Perception, & Psychophysics*, *79*(1), 333–343. doi: 10.3758/s13414-016-1206-4

Bosker, H. R. (2017b). How our own speech rate influences our perception of others. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*, 1225–1238. doi: 10.1037/xlm0000381

Bosker, H. R., Reinisch, E., & Sjerps, M. J. (2017). Cognitive load makes speech sound fast, but does not modulate acoustic context effects. *Journal of Memory and Language*, *94*, 166–176. doi: 10.1016/j.jml.2016.12.002

Bybee, J. (2006). *Frequency of use and the organization of language.* Oxford: Oxford University Press.

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, *108*(3), 804–809.

Cooper, A., Brouwer, S., & Bradlow, A. R. (2015). Interdependent processing and encoding of speech and concurrent background noise. *Attention, Perception, & Psychophysics*, *77*(4), 1342–1357.

Creel, S. C., Aslin, R. N., & Tanenhaus, M. K. (2012). Word learning under adverse listening conditions: Context-specific recognition. *Language and Cognitive Processes*, *27*(7-8), 1021–1038.

Creel, S. C., & Bregman, M. R. (2011). How talker identity relates to language processing. *Language and Linguistics Compass*, *5*(5), 190–204.

de Brouwer, A. J., Smeets, J. B., & Plaisier, M. A. (2016). How heavy is an illusory length? *i-Perception*, *7*(5), 1–5.

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

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Attention, Perception, & Psychophysics*, *67*(2), 224–238.

Goldinger, S. D. (1992). *Words and voices: Implicit and explicit memory for spoken words* (dissertation).

Indiana University.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*(2), 251–279.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson & W. J. Mullennix (Eds.), *Talker variability in speech processing* (pp. 145–165). San Diego: Academic Press.

Jones, M. R., & McAuley, J. D. (2005). Time judgments in global temporal contexts. *Perception & Psychophysics*, *67*(3), 398–417.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*(2), 148–203.

Maye, J., Weiss, D. J., & Aslin, R. N. (2008). Statistical phonetic learning in infants: Facilitation and feature generalization. *Developmental Science*, *11*(1), 122–134.

McAuley, J. D., & Miller, N. S. (2007). Picking up the pace: Effects of global temporal context on sensitivity to the tempo of auditory sequences. *Perception & Psychophysics*, *69*(5), 709–718.

McQueen, J. M., Cutler, A., & Norris, D. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science*, *30*(6), 1113–1126.

Miller, J. L. (1981). Some effects of speaking rate on phonetic perception. *Phonetica*, *38*(1–3), 159–180.

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

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

Pierrehumbert, J. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In J. Bybee & P. Hopper (Eds.), *Frequency effects and the emergence of lexical structure* (pp. 137–157). Amsterdam: John Benjamins.

Pisoni, D. B. (1993). Long-term memory in speech perception: Some new findings on talker variability, speaking rate and perceptual learning. *Speech Communication*, *13*(1), 109–125.

Pitt, M. A., Szostak, C., & Dilley, L. C. (2016). Rate dependent speech processing can be speech specific:

Evidence from the perceptual disappearance of words under changes in context speech rate. *Attention, Perception, & Psychophysics*, *78*(1), 334–345.

Pufahl, A., & Samuel, A. G. (2014). How lexical is the lexicon? Evidence for integrated auditory memory representations. *Cognitive Psychology*, *70*, 1–30.

Quené, H. (2008). Multilevel modeling of between-speaker and within-speaker variation in spontaneous speech tempo. *The Journal of the Acoustical Society of America*, *123*(2), 1104–1113.

R Core Team. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Reinisch, E. (2016). Speaker-specific processing and local context information: The case of speaking rate. *Applied Psycholinguistics*, *37*(6), 1397–1415.

Reinisch, E., Jesse, A., & McQueen, J. M. (2011). Speaking rate from proximal and distal contexts is used during word segmentation. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(3), 978.

Reinisch, E., & Sjerps, M. J. (2013). The uptake of spectral and temporal cues in vowel perception is rapidly influenced by context. *Journal of Phonetics*, *41*(2), 101–116.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, *70*(1), 27–52.

Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, *7*(5), 1074–1095.

Theodore, R. M., Miller, J. L., & DeSteno, D. (2009). Individual talker differences in voice-onset-time: Contextual influences. *The Journal of the Acoustical Society of America*, *125*(6), 3974–3982.

Wade, T., & Holt, L. L. (2005). Perceptual effects of preceding nonspeech rate on temporal properties of speech categories. *Perception & Psychophysics*, *67*(6), 939–950.

	Sentence
1	[Peter fluisterde Ilse iets verkeerd in en toen hoorde] **Ilse het** tak-/taak**je** [gezegd worden]. *Peter whispered something in Ilses ear incorrectly and then Ilse heard "the twig/task" being said.*
2	[Toen Luuk mompelend iets tegen Lotte vertelde hoorde] **Lotte het** tak-/taak**je** [gezegd worden]. *When Luuk muttered something to Lotte, Lotte heard "the twig/task" being said.*
3	[Riet probeerde de notitie te ontcijferen en plots] **kon ze het** tak-/taak**je** [onderscheiden]. *Riet was trying to decipher the note and suddenly she could discern the twig/task.*
4	[Loes twijfelde over de juiste oplossing en toch streep]**te ze het** tak-/taak**je** [door op de toets]. *Loes was unsure about the correct solution and yet she crossed out the twig/task on the test.*
5	[Toen Evelien gisteren iets onnozels wilde zeggen] **heeft ze eens** stad/staat **ge**[zegd tegen Job]. *When Evelien wanted to say something silly yesterday, she said "city/state" to Job once.*
6	[Terwijl Niels rustig zijn tijdschrift stond te lezen hebben de] **heren eens** stad/staat **te**[gen hem gebruld]. *Whilst Niels was peacefully reading his magazine, the gentlemen roared "city/state" to him once.*
7	[Femke lette goed op of ze niet ging stotteren en toen] **heeft ze eens** stad/staat **te**[gen Roos gezegd]. *Femke took care not to stutter and then she said "city/state" to Roos once.*
8	[Toen Simon de oplossing even niet meer wist fluisterde] **Nienke eens** stad/staat **in** [zijn linkeroor]. *Just as Simon could no longer remember the solution, Nienke whispered "city/state" once in his left ear.*

Table 1: Two talkers were recorded producing a set of eight Dutch stimulus sentences (English paraphrase below). These sentences were composed of an /ɑ, a:/ target word, with buffers on either side of the target, and rate-manipulated context phrases (ratio 1.6 for *slow*, 1 for *neutral*, and 0.625 for *fast*). The formatting denotes [context phrase] **buffer** target **buffer** [context phrase].

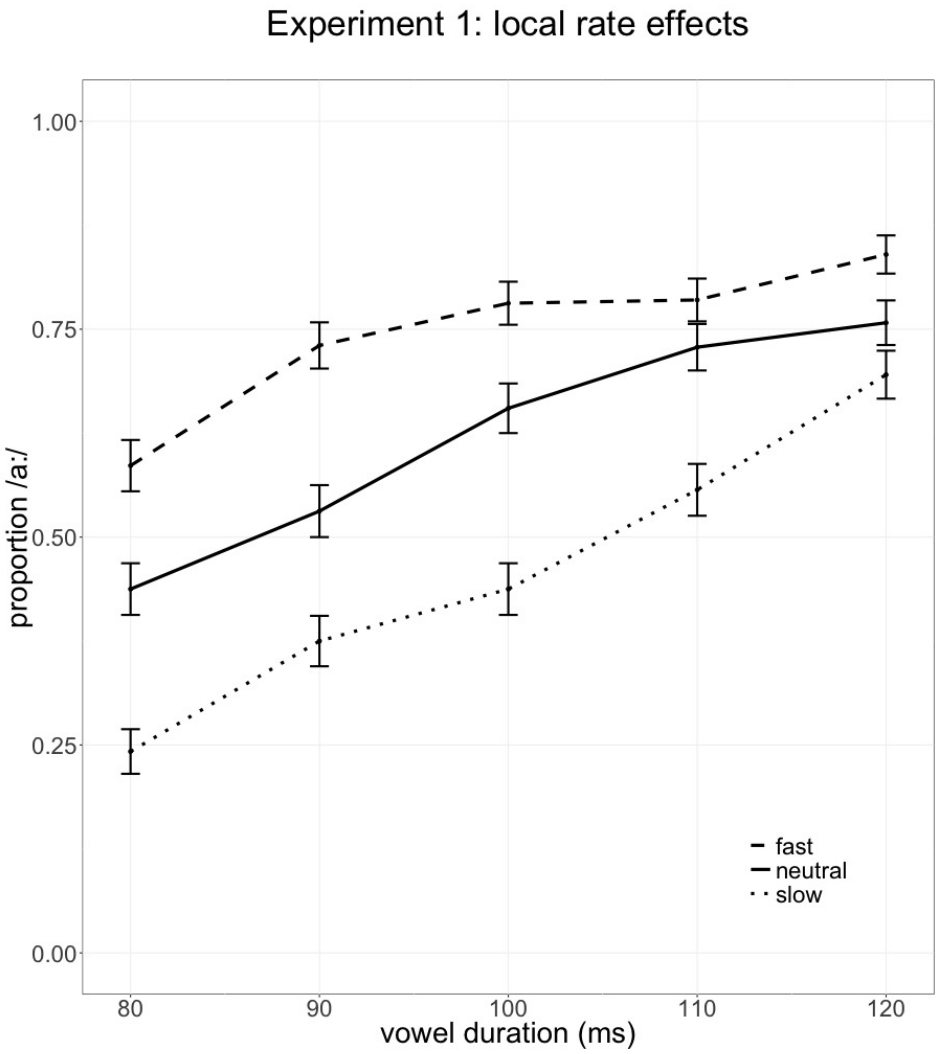

Figure 1: Average categorization data of Experiment 1 (local rate effects). The x-axis indicates Vowel Duration (80 to 120 ms). Context Rate *fast* is indicated by the dashed line, *neutral* by the solid line, and *slow* by the dotted line. Error bars represent the standard error of the mean.

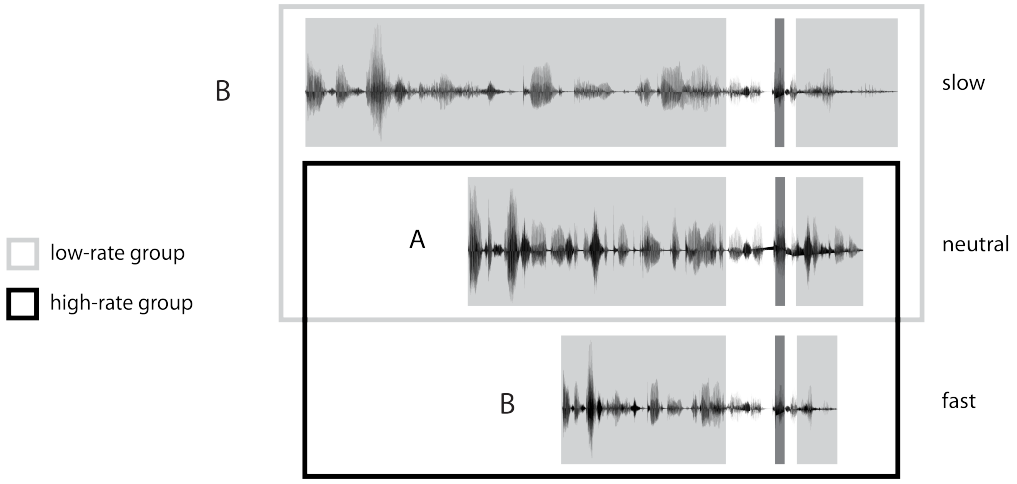

Figure 2: Experimental design of Experiment 2. Each stimulus sentence consisted of a (rate-manipulated: fast, neutral, slow) context phrase (light grey background), buffers on either side of the target (fixed duration; white background) and the target vowel itself (dark grey background). A low-rate group listened to talker B at a slow rate and talker A at a neutral rate (grey box), whereas the high-rate group listened to neutral rate from talker A, but to talker B at a fast rate (black box).

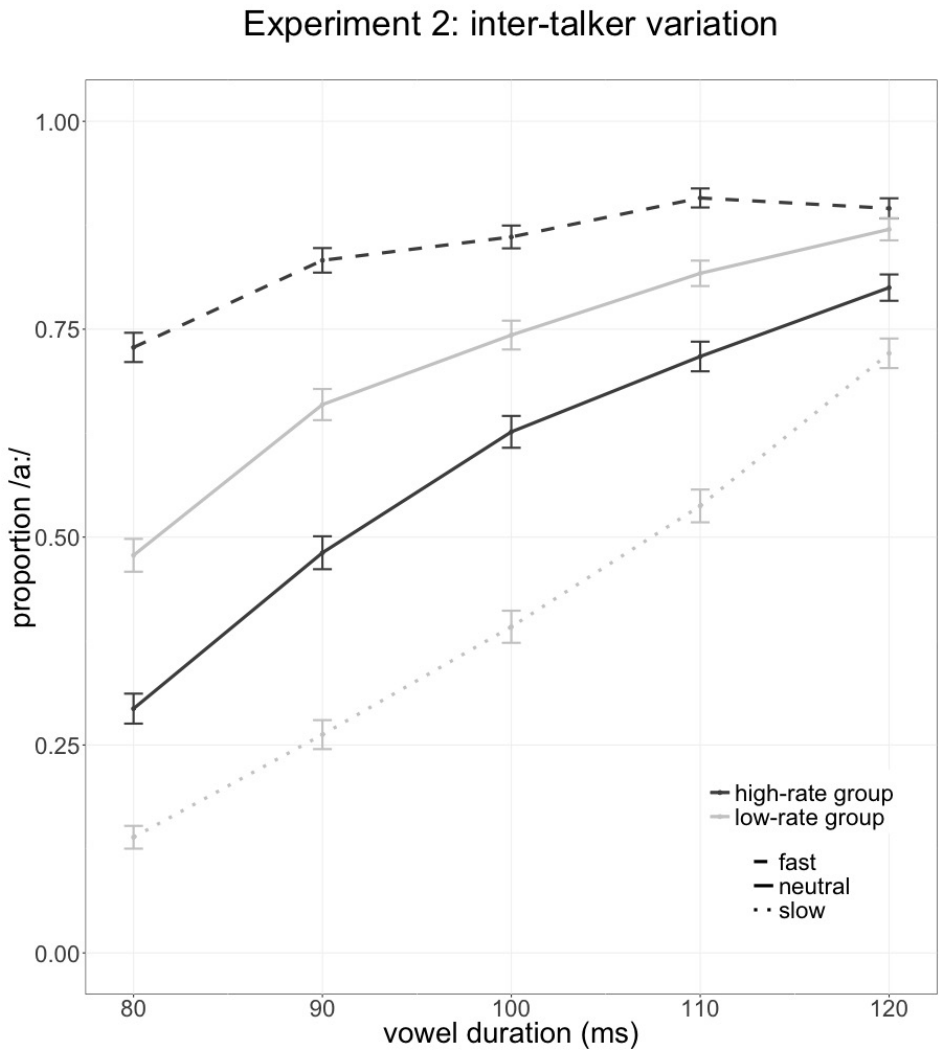

Figure 3: Average categorization data of Experiment 2 (inter-talker variation). The x-axis indicates Vowel Duration (80 to 120 ms). Rate Condition *fast* is indicated by the dashed line, *neutral* by the solid line, and *slow* by the dotted line. Colors indicate Group, with the high-rate group shown in dark grey and the low-rate group shown in light grey. Error bars represent the standard error of the mean.

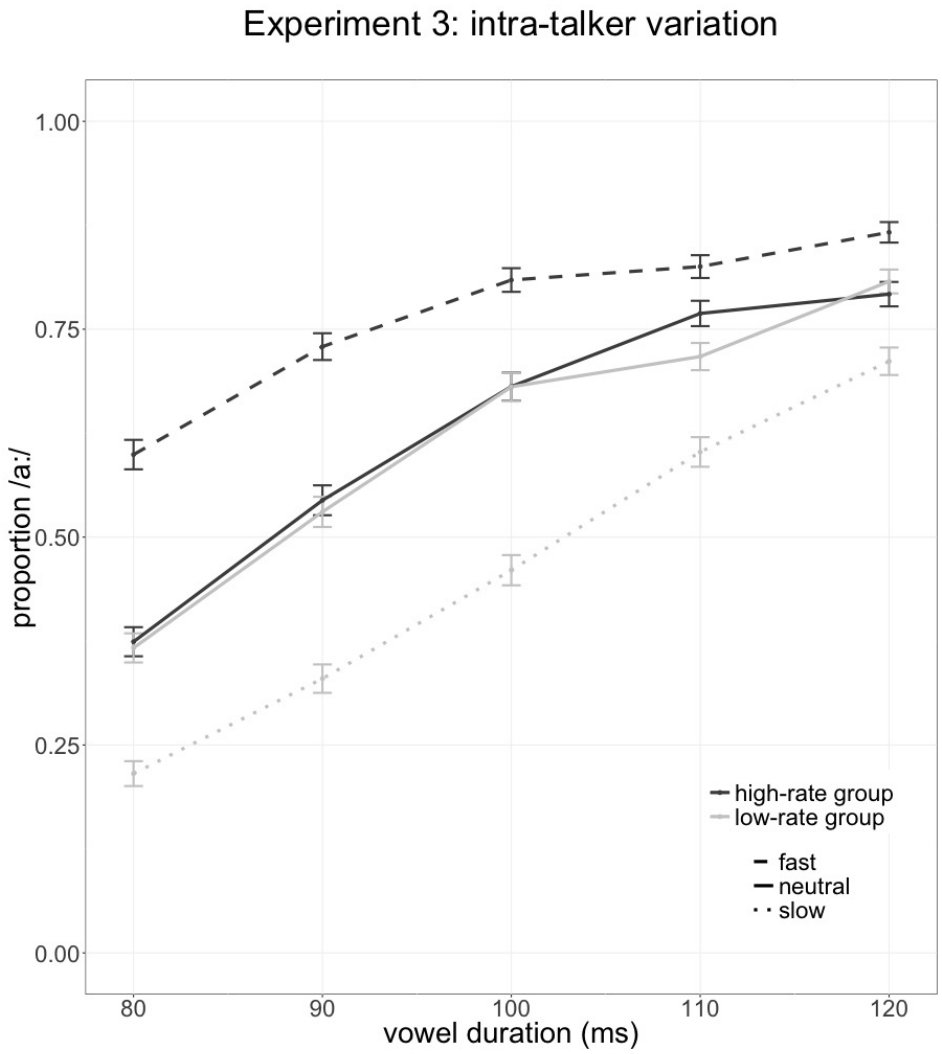

Figure 4: Average categorization data of Experiment 3 (intra-talker variation). The x-axis indicates Vowel Duration (80 to 120 ms). Rate Condition *fast* is indicated by the dashed line, *neutral* by the solid line, and *slow* by the dotted line. Colors indicate Group, with the high-rate group shown in dark grey and the low-rate group shown in light grey. Error bars represent the standard error of the mean.