Listeners normalize speech for contextual speech rate even without an explicit recognition task

Merel Maslowski,[1, a)] Antje S. Meyer,[1] and Hans Rutger Bosker[1]

Max Planck Institute for Psycholinguistics, P.O. Box 310, 6500 AH, Nijmegen, The Netherlands

Speech can be produced at different rates. Listeners take this rate variation into account by normalizing vowel duration for contextual speech rate: An ambiguous Dutch word /m?t/ is perceived as short /mɑt/ when embedded in a slow context, but long /ma:t/ in a fast context. Whilst some have argued that this rate normalization involves low-level automatic perceptual processing, there is also evidence that it arises at higher-level cognitive processing stages, such as decision making. Prior research on rate-dependent speech perception has only used explicit recognition tasks to investigate the phenomenon, involving both perceptual processing and decision making. This study tested whether speech rate normalization can be observed without explicit decision making, using a cross-modal repetition priming paradigm. Results show that a fast precursor sentence makes an embedded ambiguous prime (/m?t/) sound (implicitly) more /a:/-like, facilitating lexical access to the long target word "maat" in a (explicit) lexical decision task. This result suggests that rate normalization is automatic, taking place even in the absence of an explicit recognition task. Thus, rate normalization is placed within the realm of everyday spoken conversation, where explicit categorization of ambiguous sounds is rare.

Keywords: speech rate; rate normalization; lexical access; word recognition

[a] Merel.Maslowski@mpi.nl

I. INTRODUCTION

A key feature of speaking style is speech rate: Speech rate differs considerably across gender, age, dialect, and discourse context, but speech rate variation also occurs substantially within individual speakers and their utterances (Jacewicz *et al.*, 2010; Quené, 2008). As a result, a phonologically long vowel produced at a fast rate may have the same phonetic duration as a phonologically short vowel produced at a slow rate. The fact that talkers vary their speech rates may thus pose problems for listeners who have to distill lexical representations from the multiplicity of temporal acoustic cues. Therefore, speech rate variability may have consequences for phonological decoding, which in turn influences higher-level linguistic processes, such as lexical access and message understanding. Here, we investigated whether and how the process of rate-dependent speech perception influences lexical access.

In speech production, segment durations are shorter in fast contexts than in slow contexts. Listeners have been suggested to cope with temporal variation in the speech signal by normalizing segmental durations for surrounding speech rates (Bosker, 2017a; Diehl *et al.*, 1980; Miller, 1981).[1] In Dutch, for instance, the category boundary between a short vowel /ɑ/ (as in "mat" /mɑt/ *mat*) and a long vowel /a:/ (as in "maat" /ma:t/ *size*) can be shifted by changing the rate of a surrounding sentence context (Reinisch *et al.*, 2011; Reinisch and Sjerps, 2013). A fast speech rate typically biases target perception towards the longer category, and a slow speech rate towards the shorter category. Likewise, speech rate contexts may induce shifts in perception of other duration-cued contrasts, such as formant transitions (shift between /b/ and /w/; see Miller and Baer, 1983), voicing contrasts (e.g., shift between

/b/ and /p/; Gordon, 1988; Summerfield, 1981), singleton-geminate contrasts (Mitterer, 2018), word segmentation (Pickett and Decker, 1960; Reinisch *et al.*, 2011), and reduced word forms (Baese-Berk *et al.*, 2014; Dilley and Pitt, 2010; Pitt *et al.*, 2016). Consequently, the speech context may influence how temporally ambiguous cues embedded in this context are perceived, in turn affecting which word – for instance, a word with a long or with a short vowel – a listener hears.

Although the effect of surrounding speech rate on segmental duration perception is well established, less is known about the origin of the effect. Some have argued that rate normalization involves low-level automatic perceptual mechanisms. For instance, Reinisch and Sjerps (2013) investigated at which time point participants' vowel perception was influenced by context speech rate, using an eye-tracking paradigm. Dutch participants listened to fast and slow sentences containing minimal word pairs with a temporally and spectrally ambiguous vowel between Dutch /ɑ/ and /a:/. The authors found that listeners relied on the duration and quality of the vowel itself, as well as on rate cues in the context. Importantly, context rate modulated the uptake of vowel-internal cues immediately upon presentation of vowel onset. Toscano and McMurray (2015), also using eye-tracking, investigated effects of (preceding) contextual speech rate and (following) vowel length on perception of voice onset time (VOT) in a four-alternative forced choice task. Similar to Reinisch and Sjerps, they found that listeners relied on both speech rate and vowel-internal cues as soon as these cues were available. As such, speech rate modulated perception of VOT, whereas vowel cues, which followed the VOT contrast, were used later. Recently, evidence for the automaticity of rate normalization was found in a third eye-tracking study (Kaufeld *et al.*, in press). Kaufeld et al. compared effects of knowledge-based (morphosyntactic gender marking) and signal-based (speech rate) cues in a 2AFC task, while also measuring participants' eye movements. They found that rate normalization immediately influenced perception, even in participants with a strong behavioral preference for the knowledge-based cue. Each of these three eye-tracking studies support that speech rate effects arise early in perceptual processing.

Moreover, there is evidence that rate effects involve general auditory mechanisms, such as durational contrast (Wade and Holt, 2005) and sustained neural entrainment (Kösem *et al.*, 2018) that operate automatically, independent from attention. Bosker *et al.* (2017) recently showed that rate-dependent speech perception is unaffected by the cognitive load imposed by a non-linguistic dual-task. Rate normalization is furthermore induced by talker-incongruent contexts: A speech context from Talker A can influence perception of a target produced by Talker B (Bosker, 2017b; Maslowski *et al.*, 2018, 2019; Newman and Sawusch, 2009). These findings suggests that rate normalization happens before attentional modulation and talker segregation.

However, other studies have found evidence that effects of surrounding speech rates are dependent on which language is being spoken (with foreign languages sounding faster, inducing more 'long' responses; Bosker and Reinisch, 2017), talker identity (habitually fast talkers induce more 'long' responses; Bosker and Reinisch, 2015; Maslowski *et al.*, 2018, 2019; Reinisch, 2016), and whether or not the context sentences are intelligible (Pitt *et al.*, 2016). For instance, Pitt et al. observed that slow sine-wave speech only made following reduced function words perceptually disappear if the sine-wave speech was intelligible to the listener. These results seem to argue against an early automatic mechanism at the perceptual level. Rather, speech rate normalization in these studies seems to involve higher-level adjustments (based on who is talking or what language is being used) or lexical feedback (i.e., the important role of intelligibility of context sentences), possibly taking place at a later decision-making level.

To date, studies on rate normalization have used only a few perception tasks that all require categorization or identification. Typically, a two-alternative forced choice (2AFC) task is used, in which participants categorize an ambiguous segment embedded in a precursor as belonging to one phonemic category or another (e.g., categorizing a Dutch ambiguous /m?t/ embedded in a fast or slow context as either "mat" or "maat"; Bosker, 2017a; Reinisch *et al.*, 2011; Reinisch and Sjerps, 2013). Other studies focusing on rate-dependent perception of reduced word forms by Dilley and Pitt (2010) and Baese-Berk *et al.* (2014) have typically used transcription tasks, in which participants are presented with a written version of all speech up to an ambiguous stretch of speech and are then asked to continue the sentence. A small number of studies have used word monitoring (Baese-Berk *et al.*, 2019), transcription of entire sentences (Heffner*et al.*, 2015), or Likert scales (Miller, 1994), which also involve identification of temporally ambiguous stretches of speech. Crucially, in all these types of tasks (1) explicit attention is directed to a temporally ambiguous stretch of speech and (2) a decision is required as to what was heard. Even in eye-tracking studies (Kaufeld *et al.*, in press; Reinisch and Sjerps, 2013; Toscano and McMurray, 2015), although assessing processing in a time window before explicit categorization, attention is drawn to the ambiguous target word. Hence, both automatic and decision processes contribute to performance, making it hard to disentangle contributions from one level or the other.

Therefore, this study investigated whether rate normalization occurs when no explicit categorization is requested about the spoken ambiguous target words. By means of a cross-modal repetition priming paradigm we tested implicit consequences of speech rate processing on higher-level processes, namely lexical access. Specifically, we assessed whether ambiguous auditory primes were normalized for surrounding speech rate, in turn influencing lexical access of a following visual target word. This cross-modal priming task differs considerably from the previously used categorization and identification tasks, which require explicit decisions about the ambiguous targets. It brings us one step closer towards everyday perception of ambiguous words, where such explicit decisions are not usually made. If speech rate normalization influences cross-modal repetition priming, we can conclude that at least part of the processes responsible for rate normalization operate at an automatic processing level, independent from later decision making.

We addressed the hypothesis that speech rate cues (fast vs. slow) influence lexical access, using a cross-modal repetition priming paradigm with a lexical decision task. Repetition priming involves facilitation of the recognition of a target word when it is preceded by a prime word that is identical to the target (compared to a non-identical word) and is typically measured in response speed. In our cross-modal repetition paradigm, participants were presented with a fixed auditory context sentence containing a prime word (e.g., "Ik heb zojuist het gegeven woordje /mɑt/ gezegd" *I just said the given word /mat/*), after which they had to decide whether a string of letters (e.g., "zon", *sun*), presented visually on a computer screen, constituted a word or a non-word (see the top panel of Figure 1). Lexical decision tasks require lexical access to the orthographic string (Monsell *et al.*, 1992). As such, priming effects from preceding auditory words on lexical decision of a following target may be interpreted as influences arising from facilitation of lexical access (Marslen-Wilson and Zwitserlood, 1989). The lexical decision task is a meta-linguistic task, but the task concerns the target, not the prime. No explicit decision about the prime is required, which in our case was the ambiguous word of interest.

A set of three experiments was designed to investigate whether the rate of the precursor sentence and the spectral quality of the vowel of the prime word affect target processing. Before testing the prediction that both context rate and vowel-internal cues in the prime influence perceptual processing in an implicit task in Experiment 3, we validated the paradigm and materials in two separate experiments.

Experiment 1 validated the lexical decision paradigm with our set of stimulus words. Participants heard Dutch canonical (i.e., unambiguous) prime words embedded in a fixed precursor sentence. A written target was either identical, phonologically related, or unrelated to an auditory prime. We expected an effect of identity priming, such that responses would be faster for targets identical to their primes than for non-identical primes (Forbach *et al.*, 1974; Forster and Davis, 1984; Scarborough *et al.*, 1977). This hypothesis was confirmed. Experiment 2 then validated our stimulus set, this time using ambiguous /ɑ, aː/ words, embedded in rate-manipulated sentences (fast vs. slow) with a 2AFC task, as typically used in rate normalization studies. We predicted that a fast sentence would bias perception toward hearing a temporally and spectrally ambiguous /ɑ–aː/ vowel as long (i.e., /aː/),

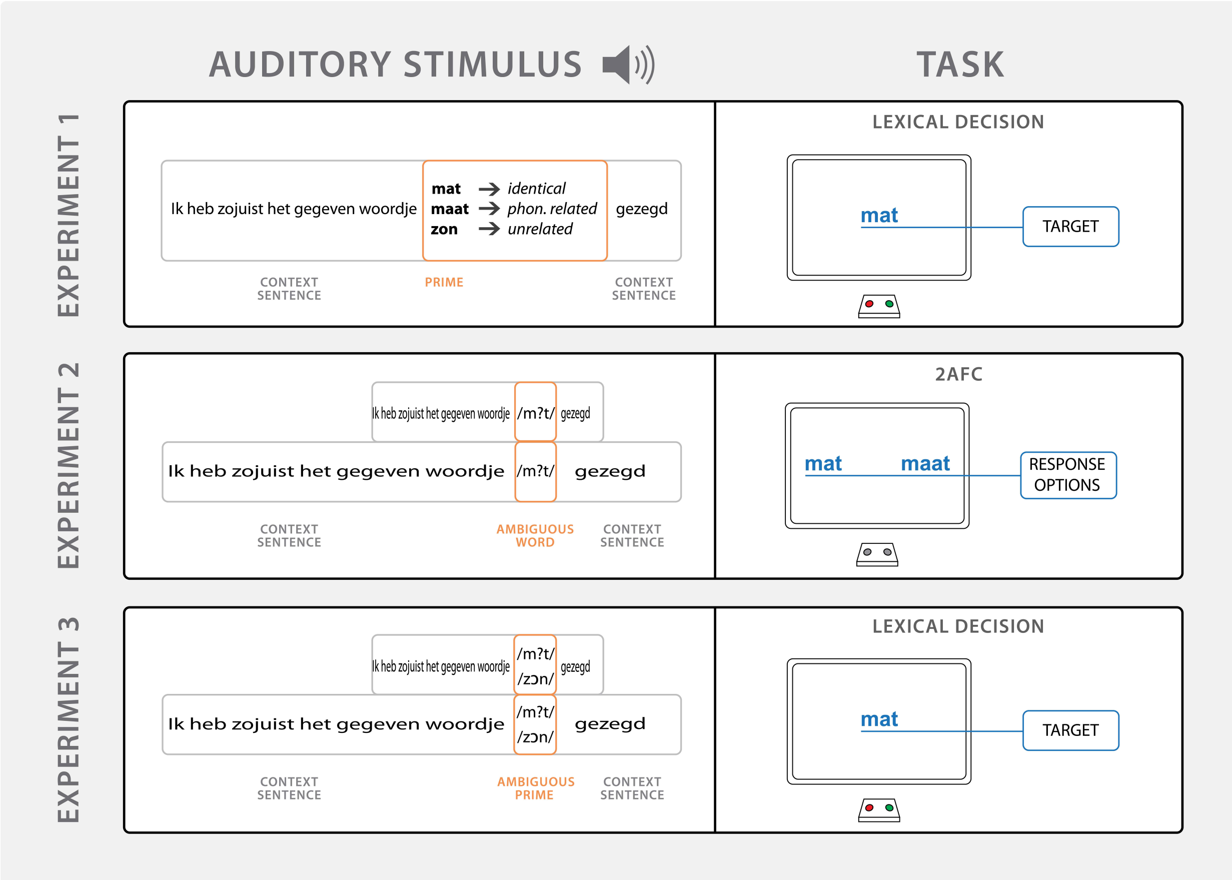

FIG. 1. Experimental design of Experiments 1–3. Experiment 1 involved a cross-modal repetition priming paradigm with a lexical decision task. Auditory primes were either identical, phonologically related, or unrelated to the following orthographic target words. Experiment 2 tested rate normalization in a two-alternative forced choice (2AFC) task. Auditory stimuli consisted of spectrally ambiguous Dutch /ɑ, aː/ vowels embedded in fast and slow context sentences. Experiment 3 combined the methods of Experiment 1 and 2, testing rate normalization of ambiguous primes with a lexical decision task.

whereas a slow sentence would bias perception towards hearing a short vowel (i.e., /ɑ/). This hypothesis was also borne out by the results.

Experiment 3 was the main experiment that combined the methods of the two previous experiments, testing rate normalization using a cross-modal repetition priming paradigm. We predicted that rate normalization should influence linguistic processing when no overt categorization response on the prime was required, supporting rate normalization as involving automatic perceptual processes. Specifically, we expected an interaction between speech rate of the prime (fast vs. slow) and the target word on the screen.

II. EXPERIMENT 1: CROSS-MODAL REPETITION PRIMING

Experiment 1 evaluated cross-modal repetition priming in a lexical decision task, testing the effect of an auditory prime on response speed to an orthographic target. Firstly, Experiment 1 aimed at validating the constructed stimuli for finding differences in reaction times in phonologically related pairs. Secondly, the experiment gives an indication of the magnitude of the differences between experimental conditions when no speech rate manipulation is performed, forming a reference for response speed differences in subsequent experiments.

A. Methods

1. Participants

Twelve native Dutch participants (female = 9, M_{age} = 22 years) without hearing or reading deficits were recruited from the Max Planck Institute participant pool. All participants gave their informed consent to participate in the experiment, as approved by the

Ethics Committee of the Social Sciences department of Radboud University (project code: ECSW2014-1003-196).

2. Design and materials

A native Dutch female talker was recorded producing each of 540 monosyllabic primes in the precursor "Ik heb zojuist het gegeven woordje [prime] gezegd" (*I just said the given word [prime]*). Creaky-voiced precursors were replaced with different recordings to facilitate digital rate-manipulation in the two following experiments. A precursor consisting of both a long pre-carrier (up to the prime word) and a short post-carrier (after the prime word) was chosen for two reasons. On the one hand, rate-manipulated stretches of speech on both sides of an acoustically ambiguous prime increases the opportunity for observing an effect of speech rate in subsequent rate-dependent speech perception experiments. On the other hand, it is desirable to keep the interval between prime and target as short as possible, in order to find an effect of repetition priming. Here, the pre-carriers had a mean duration of 1.914 s ($sd = 0.058$), and the post-carriers had a mean duration of 0.665 ($sd = 0.040$).

There were three experimental conditions, referring to three different relationships between primes and targets. Prime and target could be identical pairs (e.g., prime /mɑt/ *mat* and target "mat" *mat*), phonologically related (e.g., prime /ma:t/ *size* and target "mat" *mat*), or phonologically and semantically unrelated (e.g., prime /zɔn/ *sun* and target "mat" *mat*). Unrelated primes were monosyllabic, consisted of maximally six letters, and contained no instances of the vowels /ɑ/ and /a:/. Furthermore, they matched the target words in word frequency and dominant part-of-speech, both of which properties were extracted from

SUBTLEX-NL (Keuleers *et al.*, 2010). In total, there were 90 /ɑ, a:/ minimal pairs that were matched with an unrelated prime with the properties described above (see Supplementary materials). Similarly, there were 180 filler trials with non-word targets. Filler primes either contained an /a:/ (1/3), an /ɑ/ (1/3), or a different vowel (1/3), corresponding to the experimental trials. Filler target words always contained an /a:/ (1/2) or an /ɑ/ (1/2), as experimental target words also always contained either an /a:/ (1/2) or an /ɑ/ (1/2).

3. Procedure

The presentation of stimuli was controlled by Presentation software (v16.5; Neurobehavioral Systems, Albany, CA, USA). At trial onset, an auditory stimulus was presented through headphones, whilst a fixation point was shown on the computer screen in front of the participant. Immediately after stimulus offset, this screen was replaced with another screen with a string of letters (i.e., there was no delay between sentence offset and target onset). Participants had to indicate with a button press whether the string of letters formed a Dutch word or a non-word. If no response was given within 2 seconds after stimulus offset, a missing response was recorded. Therefore, no extreme outliers were present in the data.

The 180 experimental target words occurred once in each of three participant groups, albeit in different experimental conditions (*identical*, *phonologically related*, and *unrelated*). For the full set of 90 minimal pairs, each participant from each group responded to each combination of experimental condition and vowel 15 times. Stimulus presentation was randomized, except that for each minimal pair, one member was presented as a target in the first half of the experiment and the other member in the second half of the experiment. Which member was presented in which half was counterbalanced across participants, as were the button positions of the two response options.

The experiment started with eight practice trials with eight primes and targets without /ɑ, aː/ to familiarize participants with the paradigm. Participants were instructed to respond as fast and accurately as possible. After that, participants responded to 360 experimental trials in total, half of which were fillers. They were allowed a short break after every 36 trials. One experimental session lasted for approximately 40 minutes.

B. Results and Discussion

All participants performed above 85% in the lexical decision task, with a mean of 89.81% accuracy on words, a mean of 97.31% on non-words, and a mean of 93.56% overall[2]. Figure 2 summarizes the RTs for correct responses in each of the three experimental conditions (identical, phonologically related, and unrelated). The figure suggests that participants responded earlier to targets that were identical to their primes than to targets that were phonologically related or unrelated.

The RTs of accurate experimental trials (10.19% incorrect experimental trials excluded) were tested using a Generalized Linear Mixed Model (GLMM) from the `lme4` package (Bates *et al.*, 2015) in `R` (R Core Team, 2014). The predictors in the model were Prime Condition (categorical predictor; intercept is phonologically related) and Word Frequency (log-transformed continuous predictor). We always started with a maximal random effects structure, as recommended Barr *et al.* (2013), unless the full model failed to reach convergence. If random slopes had to be dropped due to convergence issues, slopes of the fixed effects were

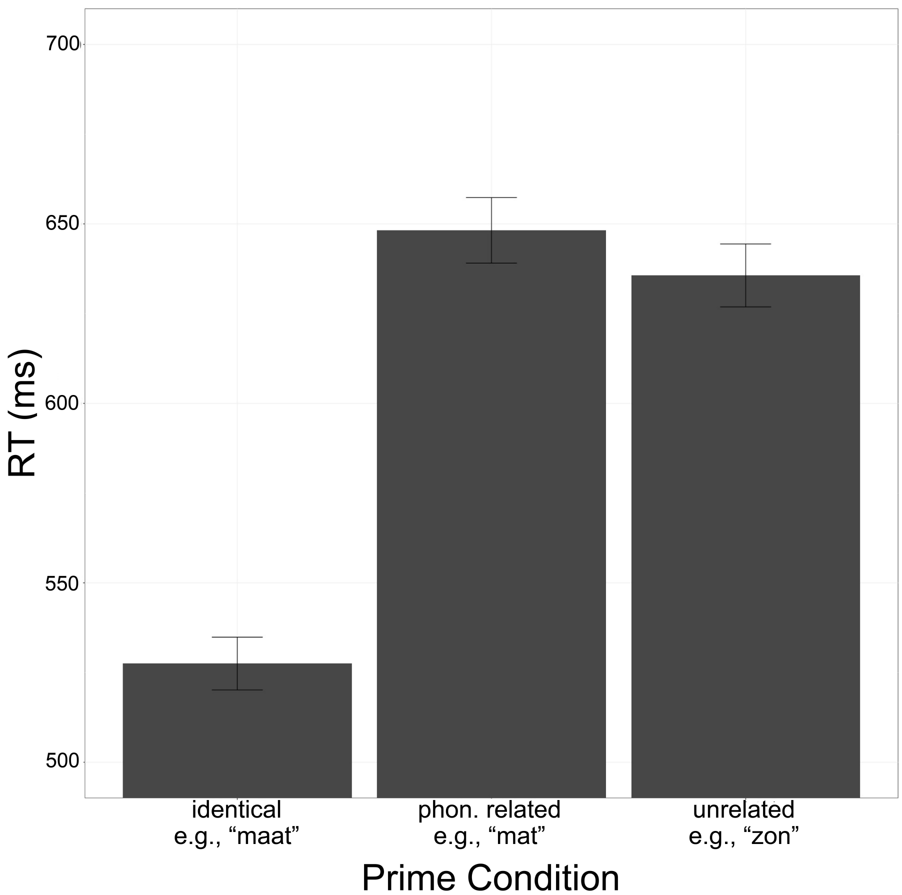

FIG. 2. Mean reaction times of Experiment 1 (cross-modal repetition priming) for correct responses in three Prime Conditions (*unrelated*, *phonologically related*, and *identical*). Error bars indicate the standard error of the mean.

gradually removed with the lowest estimated variance by both random effects (Participants and Items) simultaneously. Here, random intercepts were included for Participant nested within Group and for Target Word nested withing Minimal Pair. Random slope terms were tested for both predictors by both random factors.

Reaction times for correct responses significantly decreased when primes and targets were identical, as compared to when primes and targets were phonologically related (β =

$-106.068, t = -4.337, p = 0.001$)[3]. There was no significant difference between phonologically related and unrelated primes and targets ($\beta = -16.102, t = -0.997, p = 0.340$). Word Frequency significantly influenced reaction times ($\beta = -15.447, t = -4.713, p < 0.001$), with responses being faster to higher frequency words than to lower frequency words.

The results of the experiment indicate that responses were faster for targets identical to their primes than for phonologically related or unrelated targets. Response speed for phonologically related words was similar to the unrelated condition, which served as a baseline condition. This experiment confirms that lexical access is facilitated when a word has been primed by an identical auditory prime, replicating previous literature using similar paradigms.

III. EXPERIMENT 2: RATE NORMALIZATION IN 2AFC TASK

Experiment 2 assessed rate normalization in a 2AFC task with the same /ɑ, aː/ words as in Experiment 1. Specifically, only the auditory primes from Experiment 1 were used. This time, however, the precursor sentences surrounding the /ɑ, aː/ words were rate-manipulated (fast vs. slow), and participants categorized temporally and spectrally ambiguous /ɑ, aː/ words. That is, participants simply listened to the ambiguous tokens in fast and slow contexts and indicated which of two response options (e.g., “mat” or “maat”) they had heard (see the middle panel of Figure 1). The experiment aimed to test whether the stimulus set would elicit the typical finding that a fast context biases perception of a spectrally ambiguous /ɑ–aː/ vowel towards a long vowel /aː/, whereas a slow context biases perception of the same vowel towards hearing /ɑ/.

A. Methods

1. Participants

Fourteen native Dutch participants (female = 12; M_{age} = 24 years) recruited from the same participant pool as before gave their informed consent to participate. A priori, it was decided to exclude participants for whom the stimuli were insufficiently ambiguous (proportion of < 0.1 or > 0.9 /a:/ responses). One participant was excluded based on this criterion and another was excluded due to technical difficulties, resulting in data from 12 participants for analysis.

2. Design and materials

The same minimal pairs were used as in Experiment 1. For ten pairs used in Experiment 1, one or both members were incorrectly recognized as a non-word more than half of the time in the previous experiment. The words that were frequently identified as non-words were either very low-frequency words or verbs, and in one instance the proper noun "Saab" (automobile manufacturer). Therefore, these pairs (pairs 6, 7, 10, 13, 15, 53, 54, 56, 73, 81; see Supplementary materials) were excluded from the stimulus set of Experiment 2.

In Dutch, the vowel contrast between /ɑ/ and /a:/ is differentiated both temporally and spectrally (Adank *et al.*, 2004); /ɑ/ is shorter and has a lower F2 than /a:/. Therefore, for the remaining 80 minimal pairs, nine-step spectral continua (1: most /a:/-like; 9: most /ɑ/-like) were created in Praat (Boersma and Weenink, 2015). First, the two vowels of a minimal pair were extracted, and the durations and pitch contours of the vowels were matched (set to the mean) with PSOLA in Praat. For words with an /l/ or /r/ in coda, these segments were included as part of the vowel. Next, the vowels were linearly interpolated sample-by-sample in nine steps, with step 1 sounding most /a:/-like and step 9 sounding most /ɑ/-like.The weighted sounds of the vowel pair were mixed, such that the first step was based on (1/9*1 =) 0.11 of the /ɑ/-vowel, and (1/9*8=) 0.89 of the /a:/-vowel, the second step (1/9*2 =) 0.22 and (1/9*7=) 0.78, and so on.

The resulting spectral vowel continua were embedded in their consonantal frames and piloted in a 2AFC online pilot, in which participants (N = 20) were asked to categorize which member of a minimal pair they heard. From the results of this pilot study, three steps from the continuum of each pair were selected that were around 75% /a:/, 50% /a:/, and 25% /a:/ categorization (see Figure 3). As a result, the three selected steps for each pair were not necessarily equally spaced in acoustic distance, but rather in perceptual distance. Based on this pilot, another five minimal pairs (pairs 14, 18, 37, 46, and 68; see Supplementary materials) were excluded, as a consequence of not being perceived as sufficiently ambiguous between the two members. This resulted in a total of 75 pairs, which were then embedded in the same fixed precursor sentence as in Experiment 1. This time, the entire precursor sentence was rate-manipulated through linear expansion (factor 1.5) and linear compression (factor 0.67) using PSOLA in Praat (Boersma and Weenink, 2015), resulting in a slow and a fast precursor sentence. The precursor sentence consisted of a pre-carrier up to the prime word (fast: M = 1.282 s, sd = 0.039; slow: M = 2.871 s, sd = 0.087) and a post-carrier after the prime word (fast: M = 0.445 s, sd = 0.026; slow: M = 0.997 s, sd = 0.059). For each of the 90 minimal pairs, one of the two sentence recordings of a pair was used as

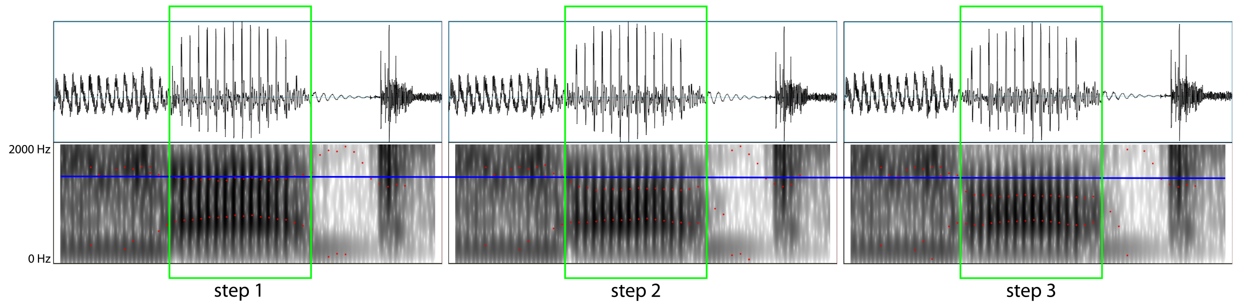

FIG. 3. Spectrograms (0 – 2000 Hz) of the three steps of the same minimal pair “hak/haak”. Step 1 is most /a:/-like (relatively high F2) and step 3 is most /ɑ/-like (relatively low F2). The green rectangles show the vowel portions. The red dots show the formant trajectories. The blue line is drawn to more easily see that F2 decreases from the left panel to the right.

the precursor sentence for that pair. Within-pair cross-splicing did occur, but because the precursor sentence and the consonantal frame of a pair was always the same, this cross-splicing was never noticeable.

Each pair was presented in six different conditions, that is, in three different spectral steps (75% /a:/, 50% /a:/, and 25% /a:/), which were embedded in two speech rate contexts (fast/slow). This resulted in 450 unique stimuli in total.

3. Procedure

Again, the Presentation software package (v16.5; Neurobehavioral Systems, Albany, CA, USA) was used to control the experiment. During presentation of each auditory stimulus, a fixation cross was shown on the screen. Immediately after stimulus offset, this screen was replaced by a different screen with two response options, each of them representing one of the members of a minimal pair on either side of the screen. Which of the two members was positioned on the right of the screen and which on the left was counterbalanced across participants. Participants were instructed to indicate which of two words they had heard in a sentence by responding with a left/right button press (corresponding to the positions of the response options on the screen) on a button box as fast and accurately as possible. They had four seconds to do so, before a missing response was recorded. The experiment started with a practice round with four fast and four slow trials to make the participant comfortable with the used speech rates. Each of the 450 stimuli were presented to each participant once and the experiment lasted for about 50 minutes.

B. Results and Discussion

The categorization data of Experiment 2 are represented in Figure 4. As expected, participants reported hearing more long /a:/ words when vowels were spectrally more /a:/-like (lower steps on the vowel continua), and fewer long vowels when they were more /ɑ/-like (higher steps on continua). The difference between the two lines indicates that participants also reported hearing more long vowels in fast rate contexts than in slow contexts.

The binomial categorization responses (/ɑ/ responses coded as 0; /a:/ responses coded as 1) of Experiment 2 (0 missing responses) were tested with a GLMM with a logistic linking function to analyze whether the current stimuli generated the typical finding that a fast speech rate context leads to more /a:/ responses than a slow context. The model included fixed effects for Vowel Step (continuous predictor; centered and divided by one standard deviation), Rate Condition (categorical predictor; intercept is fast), and their interaction. The full random effect structure was used, with intercepts for Participant and Minimal Pair

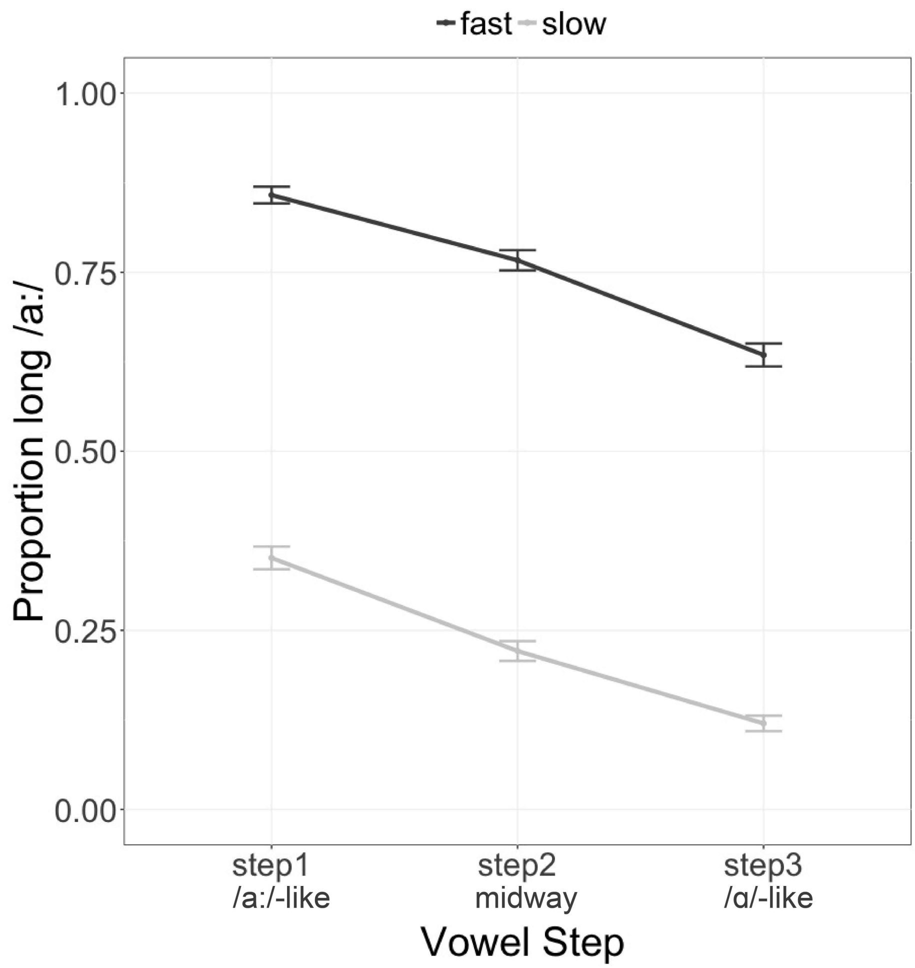

FIG. 4. Average categorization data of Experiment 2 (rate normalization in 2AFC task). The x-axis indicates Vowel Step (1: /a:/-like; 3: /ɑ/-like). Colours indicate Rate Condition, with the *fast* condition shown in dark grey and the *slow* condition shown in light grey. Error bars indicate the standard error of the mean.

and random slopes for Vowel Step, Rate Condition, and their interaction by both random effects.

The proportion of long /a:/ responses significantly decreased with Vowel Step ($\beta = -0.711, z = -8.900, p < 0.001$), indicating that spectrally more /ɑ/-like vowels were less often categorized as a long /a:/ than spectrally more /a:/-like vowels. Moreover, the proportion of /a:/ responses also significantly decreased for the slow Rate Condition ($\beta = -3.556, z = -15.576, p < 0.001$) relative to the fast condition mapped onto the intercept. This result indicates that speech rate context modulated perception of the target vowel. The interaction between Vowel Step and Rate Condition was not significant ($\beta = -0.121, z = -1.135, p = 0.256$).

As expected, categorization data revealed effects of the spectral continua and of the precursor, with fast precursors biasing perception towards /a:/. As such, the experiment replicates rate normalization effects observed previously in studies using a similar 2AFC design (Bosker, 2017a; Kaufeld *et al.*, in press; Reinisch and Sjerps, 2013).

IV. EXPERIMENT 3: RATE NORMALIZATION IN REPETITION PRIMING

Experiment 3 involved cross-modal repetition priming with a lexical decision task, combining the methods of the previous experiments. That is, the rate-manipulated precursors with spectrally ambiguous /ɑ, a:/ words from Experiment 2 were used as primes to test RTs on the same orthographic targets as in Experiment 1 (see bottom the panel of Figure 1). This experiment tested whether speech rate effects are induced even when no explicit attention is drawn to the spectrally ambiguous word.

A. Methods

1. Participants

Eighty native Dutch participants (female = 55; $M_{age} = 22$ years) were recruited from the participant pool of the Max Planck Institute and gave their consent to participation.

2. Design and materials

The materials included the rate-manipulated stimuli with spectrally ambiguous vowels from Experiment 2 as primes and the target items (words and non-words) from Experiment 1 as target words (minus the 15 excluded pairs). Additionally, Experiment 3 contained the control primes of Experiment 1, that is, the unrelated words without the /ɑ–a:/ contrast. For consistency, control prime precursors were also rate-manipulated. Each minimal pair appeared as two targets (e.g., V "mat" and V "maat") with four primes (unrelated; step 1: 75% /a:/; step 2: 50% /a:/; step 3: 25% /a:/), all combined with a fast and a slow precursor. This resulted in a stimulus set of 1200 unique test stimuli (75 minimal pairs x 2 targets x 4 primes x 2 rates).

3. Procedure

The experimental task was identical to that of Experiment 1. Eight lists consisting of 150 different test trials (and with each target appearing only once in every list) were constructed using a Latin square design. In every list, one member of a minimal pair appeared as a target in the first half of the experiment and the other in the second half. The 75 test stimuli within each half were presented in randomized order together with equally many filler trials with non-word targets, resulting in 300 trials in total. Stimulus presentation was identical to the procedure in Experiment 1. One experimental session lasted for about 35 minutes.

B. Results and Discussion

All participants performed above 85% accuracy in the lexical decision task, with a mean of 93.88% on words, a mean of 97.76% on non-words, and 95.82% overall. Figure 5 summarizes the reaction times (RTs) for the correct responses in four prime conditions (including the control condition *unrelated primes*). The top panel shows that RTs are shorter with a matching /a:/-like vowel in the prime (step 1) than a vowel midway between /a:/ and /ɑ/ (step 2) or an /ɑ/-like vowel (step 3). This is consistent with the identical versus different contrast in Experiment 1. Moreover, for each prime, we observed a rate normalization effect: RTs were shorter for fast precursors sentences (making the prime appear longer) than slow sentences preceding long targets. For short targets (bottom panel), the opposite pattern is seen: RTs were longer for fast precursors than for slow precursors, in which the prime sounds shorter.

The RTs on trials with an "a" or "aa" target (e.g., "mat" and "maat"; i.e., excluding control trials with "zon" as target) were tested with a Linear Mixed Model from the `lme4` package (Bates *et al.*, 2015) in `R` (R Core Team, 2014). The fixed factors in the model included Target Word (long vs. short; categorical predictor; sum-to-zero coded), Prime Condition (vowel step 1 to 3 as a continuous predictor; centered and divided by one standard deviation), Precursor Rate (categorical predictor; sum-to-zero coded), two-way interactions between these three predictors, as well as a three-way interaction. Note that the unrelated primes (that served as a control condition) were excluded from analysis to treat Prime Condition as a continuous variable. The random effect structure consisted of Participant nested within Group and Item nested within Minimal Pair.

RTs significantly increased for Target Word ($\beta = 26.459, t = 2.356, p = 0.020$)[3], with longer RTs for the long members of minimal pairs than for the short members of the pairs. This result may be expected given that longer words (with two vowel characters; "aa") take longer to read than shorter words (with one vowel character; "a"). RTs were also significantly affected by Prime Condition ($\beta = 5.514, t = 2.776, p = 0.006$); RTs were longer for more /ɑ/-like vowels than for /a:/-like vowels, perhaps because /ɑ/-words generally have higher neighborhood densities than /a:/-words (Marian *et al.*, 2012). Precursor Rate was not significant ($\beta = 2.528, t = 0.637, p = 0.524$), showing no overall main effect of speech rate context. The model showed a significant interaction between Target Word and Prime Condition ($\beta = 29.087, t = 7.320, p < 0.001$), indicating shorter RTs for long targets with more /a:/-like primes, but longer for short targets with more /a:/-like primes. The interaction between Target Word and Precursor Rate was also significant ($\beta = -83.641, t = -10.529, p < 0.001$). This interaction indicates that RTs were shorter for long targets with fast primes, but longer RTs for the same long targets with slow primes (and vice versa for short targets). The interaction between Prime Condition and Precursor Rate was not significant ($\beta = -4.671, t = -1.176, p = 0.239$), nor was the three-way interaction between all predictors ($\beta = 3.624, t = 0.458, p = 0.646$).

These results demonstrate that RTs were longer when there was a mismatch between Target Word and Precursor Rate. A fast precursor followed by a long target led to faster responses than the same target word after a slow prime. This result replicates previously reported rate normalization effects with a lexical decision task where no explicit attention is drawn to the spectrally ambiguous word in the prime.

V. GENERAL DISCUSSION

This study investigated effects of rate normalization on the speed of word recognition. Previous studies have typically studied the phenomenon of speech rate normalization with explicit tasks, in which participants' attention is drawn directly to a temporally ambiguous stretch of speech, after which they are asked to make a decision about what they have heard – something relatively long (e.g., /aː/ rather than /ɑ/; Reinisch and Sjerps, 2013) or something relatively short (/ɑ/). However, such tasks cannot distinguish between processes happening at an automatic processing level and those happening at a later decision-making level when a response is required. In the present study, we investigated whether rate normalization is in fact as automatic as argued by, for instance, Wade and Holt (2005) and Bosker *et al.* (2017), by assessing whether rate normalization can be observed outside the typical explicit recognition tasks.

A set of three experiments was conducted to test consequences of rate normalization on lexical access by means of a cross-modal repetition priming paradigm. The first two experiments involved basic paradigms for cross-modal repetition priming and speech rate normalization, testing two preconditions needed for Experiment 3. Experiment 1 validated

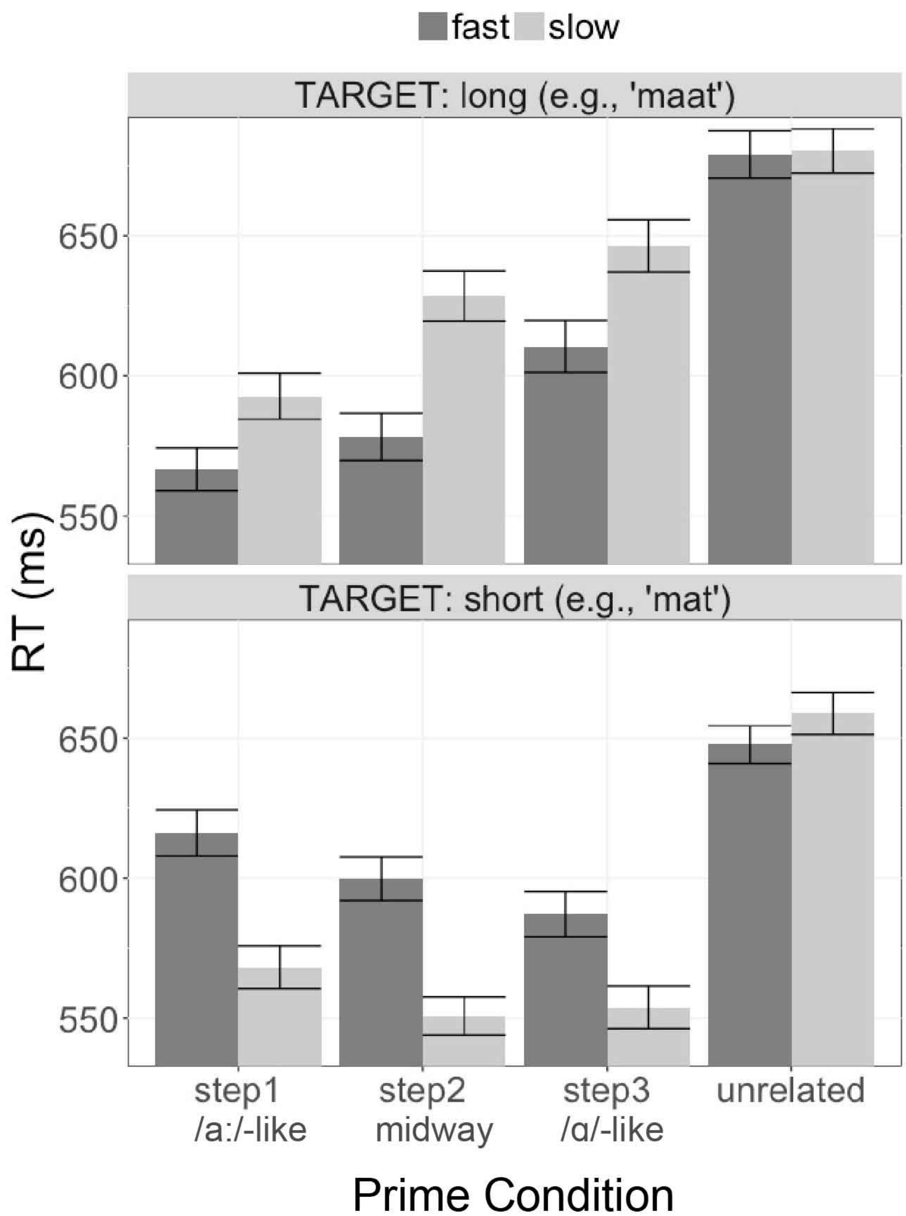

FIG. 5. Mean reaction times of Experiment 3 (rate normalization in repetition priming) for correct responses in four Prime Conditions. These conditions consisted of Vowel Step 1 (most /a:/-like), 2 (midway between /a:/ and /ɑ/), and 3 (most /ɑ/-like), as well as an unrelated control condition. Colours indicate Rate Condition, with the *fast* condition shown in dark grey and the *slow* condition shown in light grey. Error bars indicate the standard error of the mean.

the cross-modal repetition priming paradigm with our auditory primes and orthographic targets. The results of this experiment confirmed the hypothesis that lexical access of a target word is facilitated when it is identical to the prime, relative to a non-identical prime (whether or not phonologically related to the target). The second experiment showed speech rate effects with the same materials in a typical 2AFC paradigm, with fast contexts biasing participants towards hearing long vowel words, and slow contexts inducing a bias to short vowel words.

In Experiment 3, the stimuli of Experiment 2 were combined with the cross-modal repetition priming paradigm used in Experiment 1. We predicted an interaction between speech rate condition (fast/slow) and target word condition (long/short). The results of the experiment supported our prediction: When the rate of a precursor sentence was slow (biasing participants to hear /ɑ/ in the prime word), the response time to a target word with an "a" was shorter than to a target word containing "aa". Similarly, when the rate of the precursor was fast (biasing perception towards /a:/), response times to "aa" target words were shorter. These results demonstrate that speech rate normalization bears direct consequences for higher-level linguistic processing further downstream, such as lexical access.

These findings provide strong evidence for rate normalization not being task-driven. The results show that rate normalization occurs, at least in part, at an automatic processing level rather than at a later decision-making level. They corroborate earlier findings that rate normalization involves automatic perceptual mechanisms. For instance, speech rate effects have been shown to be insensitive to talker voice changes (Maslowski *et al.*, 2018, 2019; Newman and Sawusch, 2009) and they have been suggested to involve sustained neural entrainment (Kösem *et al.*, 2018). Moreover, the results of Experiment 3 strongly indicate that effects of rate normalization occur even when no explicit attention is directed to a phonologically ambiguous prime word. This finding corroborates Bosker *et al.* (2017), who showed that spectral and temporal rate normalization is unaffected by attention. It also indicates that rate normalization takes place in the absence of explicit categorization of the ambiguous segments. Listeners automatically take into account contextual speech rate when encountering temporally and spectrally ambiguous sounds. Crucially, this means that rate-dependent speech perception may be part of everyday speech processing, where no explicit categorization occurs. Although our paradigm did not require participants to respond to the primes, which were created by rate normalization, they had to perform an explicit categorization task on a different stimulus. Evidently, such tasks are rarely performed in everyday contexts. Future work may aim to replicate the paradigm without such explicit decisions.

The results of the current study may be explained by a cue integration framework. In such a framework, listeners are thought to make use of multiple cues (e.g., vowel length, vowel quality, speech rate, speaker, etc.) as soon as they are available, with more reliable cues being weighted heavier than less reliable cues (Martin, 2016; Toscano and McMurray, 2012). In our study, such a framework would predict that both vowel-internal cues (i.e., vowel condition in three steps from /a:/ to /ɑ/) as well as vowel-external contextual cues (contextual speech rate that was fast or slow) should affect perception as soon as they are presented and even outside a 2AFC paradigm. Experiment 3 showed that both of these factors influenced perceptual processing of a prime, as evidenced by shorter reaction times for target words that were perceived as identical to the prime word than for non-identical words as a consequence of either factor. These results support earlier findings by Toscano and McMurray (2015), who similarly found that speech rate and vowel quality affected speech perception independently. They interpreted their results as acoustic cues being processed directly, whereas contextual cues such as rate modulate the uptake of these acoustic cues. The results of the current study confirm that both types of cues are used independently of each other, but go beyond the study by Toscano and McMurray (2015) by using a paradigm in which no explicit decisions about ambiguous acoustic cues are required.

The evidence presented here for rate normalization arising at the level of perceptual processing leads to the question how these findings tie in with speech rate effects that seem to happen at later levels (Bosker and Reinisch, 2017; Maslowski *et al.*, 2018, 2019; Pitt *et al.*, 2016). Different effects could emerge at different levels of word recognition. That is, some rate normalization processes may take place at an obligatory perceptual level, whereas other processes may take place at a later cognitive level. Bosker *et al.* (2017) proposed a hierarchical two-stage model for temporal and spectral normalization processes that incorporates this hypothesis. They distinguish between a first stage that involves early and automatic adjustments and a second stage that involves later cognitive adjustments. They argue that, because the first stage is automatic, rate normalization of this type is not sensitive to attention and directly modulates perception. The second stage includes effects that are sensitive to signal-extrinsic indexical properties, such as talker or conversational context.

The effects of rate normalization on lexical access in this study may be interpreted as arising at the first stage of temporal normalization, in turn affecting other linguistic mechanisms such as lexical access further downstream. The effects are induced even when no explicit attention is drawn to the temporally and spectrally ambiguous word. More generally, this study stresses that in the great range of acoustic cues individuals encounter when listening to speech, they reliably take into account speech rate information in order to interpret a message.

LIST OF FIGURES

ACKNOWLEDGMENTS

We thank Johanne Tromp for lending her voice to the stimuli. We also thank Joe Toscano, Effie Kapnoula, and one anonymous reviewer for their valuable comments and suggestions.

See supplementary material at [URL] for Stimulus characteristics of Dutch minimal /ɑ, a:/ pairs in Experiment 1 to 3.

[1]This phenomenon of a shift in the phonetic category boundary between two temporally contrastive sounds due to the contextual speech rate has also been referred to as "rate-dependent speech perception" or "context compensation". In this paper, the term "rate normalization" is used for consistency with our previous papers, without making any theoretical claims about the abstractness of speech sounds.

[2]All data have been made available on: `https://osf.io/437qw/`

[3]All p-values and t-statistics were obtained from the `lmerTest` package in `R`, which provides no degrees of freedom. Note that the contribution of each predictor was also assessed by statistical comparison of a model including each predictor or interaction between predictors and a model without the predictor, using the `anova()` function in `R`. The p-values of the likelihood ratio tests were identical to those produced by `lmerTest`.

Adank, P., Van Hout, R., and 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.

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

Baese-Berk, M. M., Dilley, L. C., Henry, M. J., Vinke, L., and Banzina, E. (**2019**). "Not just a function of function words: Distal speech rate influences perception of prosodically weak syllables," Attention, Perception, & Psychophysics **81**(2), 571–589.

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (**2013**). "Random effects structure for confirmatory hypothesis testing: Keep it maximal," Journal of Memory and Language **68**(3), 255–278.

Bates, D., Mächler, M., Bolker, B., and 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., and Weenink, D. (**2015**). "Praat: doing phonetics by computer computer program. Version 5.4. 09" http://www.praat.org/.

Bosker, H. R. (**2017**a). "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. (**2017**b). "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., and Reinisch, E. (**2015**). "Normalization for speechrate in native and non-native speech," in *18th International Congress of Phonetic Sciences 2015 [ICPhS XVIII]*, International Phonetic Association.

Bosker, H. R., and Reinisch, E. (**2017**). "Foreign languages sound fast: Evidence from implicit rate normalization," Frontiers in Psychology **8**, 1063.

Bosker, H. R., Reinisch, E., and 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.

Diehl, R. L., Souther, A. F., and Convis, C. L. (**1980**). "Conditions on rate normalization in speech perception," Perception & Psychophysics **27**(5), 435–443.

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

Forbach, G. B., Stanners, R. F., and Hochhaus, L. (**1974**). "Repetition and practice effects in a lexical decision task," Memory & Cognition **2**(2), 337–339.

Forster, K. I., and Davis, C. (**1984**). "Repetition priming and frequency attenuation in lexical access," Journal of Experimental Psychology: Learning, Memory, and Cognition **10**(4), 680–698.

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

Heffner, C. C., Newman, R. S., Dilley, L. C. and Idsardi, W. J. (**2015**). "Age-related differences in speech rate perception do not necessarily entail age-related differences in speech rate use," Journal of Speech, Language, and Hearing Research **58**(4), 1341–1349.

Jacewicz, E., Fox, R. A., and Wei, L. (**2010**). "Between-speaker and within-speaker variation in speech tempo of American English," The Journal of the Acoustical Society of America **128**(2), 839–850.

Kaufeld, G., Ravenschlag, A., Meyer, A. S., Martin, A. E., and Bosker, H. R. (**in press**). "Knowledge-based and signal-based cues are weighted flexibly during spoken language comprehension.," Journal of Experimental Psychology: Learning, Memory, and Cognition.

Keuleers, E., Brysbaert, M., and New, B. (**2010**). "Subtlex-nl: A new measure for dutch word frequency based on film subtitles," Behavior Research Methods **42**(3), 643–650.

Kösem, A., Bosker, H. R., Takashima, A., Meyer, A., Jensen, O., and Hagoort, P. (**2018**). "Neural entrainment determines the words we hear," Current Biology doi: 10.1016/j.cub.2018.07.023.

Marian, V., Bartolotti, J., Chabal, S., and Shook, A. (**2012**). "Clearpond: Cross-linguistic easy-access resource for phonological and orthographic neighborhood densities," PloS one **7**(8), e43230, doi: 10.1371/journal.pone.0043230.

Marslen-Wilson, W., and Zwitserlood, P. (**1989**). "Accessing spoken words: The importance of word onsets.," Journal of Experimental Psychology: Human Perception and Performance **15**(3), 576–585, doi: 10.1037/0096-1523.15.3.576.

Martin, A. E. (**2016**). "Language processing as cue integration: Grounding the psychology of language in perception and neurophysiology," Frontiers in Psychology **7**, 1–17.

Maslowski, M., Meyer, A. S., and Bosker, H. R. (**2018**). "Listening to yourself is special: Evidence from global speech rate tracking," PloS one **13**(9), e0203571.

Maslowski, M., Meyer, A. S., and Bosker, H. R. (**2019**). "How the tracking of habitual rate influences speech perception," Journal of Experimental Psychology: Learning, Memory, and Cognition **45**(1), 128–138, doi: 10.1037/xlm0000579.

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

Miller, J. L. (**1994**). "On the internal structure of phonetic categories: A progress report," Cognition **50**(1–3), 271–285, doi: 10.1016/0010-0277.

Miller, J. L., and Baer, T. (**1983**). "Some effects of speaking rate on the production of /b/ and /w/," The Journal of the Acoustical Society of America **73**(5), 1751–1755.

Mitterer, H. (**2018**). "The singleton-geminate distinction can be rate dependent: Evidence from maltese," Laboratory Phonology: Journal of the Association for Laboratory Phonology **9**(1), 6.

Monsell, S., Patterson, K. E., Graham, A., Hughes, C. H., and Milroy, R. (**1992**). "Lexical and sublexical translation of spelling to sound: Strategic anticipation of lexical status.," Journal of Experimental Psychology: Learning, Memory, and Cognition **18**(3), 452–467, doi: 10.1037/0278-7393.18.3.452.

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

Pickett, J., and Decker, L. R. (**1960**). "Time factors in perception of a double consonant," Language and Speech **3**(1), 11–17.

Pitt, M. A., Szostak, C., and 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.

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*, R Foundation for Statistical Computing, Vienna, Austria, 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., and 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., and 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.

Scarborough, D. L., Cortese, C., and Scarborough, H. S. (**1977**). "Frequency and repetition effects in lexical memory.," Journal of Experimental Psychology: Human perception and performance **3**(1), 1–17.

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

Toscano, J. C., and McMurray, B. (**2012**). "Cue-integration and context effects in speech: Evidence against speaking-rate normalization," Attention, Perception, & Psychophysics **74**(6), 1284–1301.

Toscano, J. C., and McMurray, B. (**2015**). "The time-course of speaking rate compensation: Effects of sentential rate and vowel length on voicing judgments," Language, Cognition, and Neuroscience **30**(5), 529–543.

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